Popularnonaukowe streszczenie projektu

Projekt związany jest z problematyką projektowania efektywnych metody automatycznej klasyfikacji wzorców w przypadku danych niezbalansowanych i zadania klasyfikacji wieloklasowej. Zbiorami niezbalansowanymi nazywamy takie zbiory, w których przykłady z niektórych klas występują zdecydowanie częściej (klasy większościowe), niż przykłady z pozostałych klas (klasy mniejszościowe). Tradycyjne metody klasyfikacji wzorców są dedykowane problemom, gdzie nie występują duże różnice w częstościach pojawiania się obiektów z różnych klas, jednakże w praktycznych zastosowaniach takie przypadki są dość częste. W przypadku problemu dwuklasowego związki pomiędzy klasami są łatwe do zdefiniowania, gdyż jedna z klas jest klasą mniejszościową, a druga klasą większościową. W przypadku zadania klasyfikacji wieloklasowej relacja pomiędzy klasami nie musi być już tak oczywista, gdyż jedna z klas może mieć charakter klasy większościowej w stosunku do niektórych klas, a w stosunku do innych może być klasą mniejszościową.
Problemy związane z danymi niezbalansowanymi można znaleźć w szeregu współczesnych, praktycznych zadaniach decyzyjnych, m.in. w bankowości (problem wykrywania oszustw w transakcjach), bezpieczeństwie systemów komputerowych (filtracja poczty elektronicznej, czy projektowanie systemów detekcji intruzów) , czy też w medycynie (np. diagnostyka typów wtórnego nadciśnienia tętniczego).

W trakcie tego projektu chcemy udowodnić hipotezę, że **możliwym jest zaprojektowanie efektywnych metod klasyfikacji wieloklasowej dedykowanych zbiorom niezbalansowanym bez potrzeby dekomponowania problemu uczenia klasyfikatora na podproblemy klasyfikacji binarnej**.

Głównymi rezultatami projektu będą reguły jakimi należy się kierować w trakcie projektowania metod uczenia klasyfikatorów bazujących na danych niebalansowanych, a także nowe algorytmy klasyfikacji tego typu danych, metody wstępnej obróbki danych niezbalansowanych, a także pakiet softwarowy zawierający komputerowe implementację zaproponowanych w trakcie projektu algorytmów.
W celu oceny jakości zaproponowanych metod zostanie przeprowadzona ich szczegółowa analiza eksperymentalna z wykorzystaniem otwartych platform programowych takich jak KNIME, umożliwiające rozwój własnego oprogramowania w środowiskach Java, R, czy Matlab.
W trakcie projektu planuje się realizacje następujących zadań badawczych:

- Opracowanie metod oceny lokalnej trudności niezbalansowanego zbioru danych problemu wieloklasowego
- Propozycja nowych algorytmów umożliwiających zbalansowanie zbiorów danych dla zadania wieloklasowego
- Opracowanie modyfikacji metod klasyfikacji niezbalansowanych zbiorów dla zadań wieloklasowych
- Propozycja nowych metod klasyfikacji kombinowanej dedykowanych niezbalansowanym zbiorom dla zadań klasyfikacji niezbalansowanej
- Implementacja i eksperymentalna ocena opracowanych metod dedykowanych niezbalansowanym zbiorom dla zadań klasyfikacji wieloklasowej.

Wykorzystanie wspomnianych metod w analizie danych niezbalansowanych dla problemów klasyfikacji wieloklasowej wydaje się obecnie słabo dostrzegane, stąd celem projektu jest wypełnienie tej luki. Opracowane metody mogą zatem być wykorzystane przez firmy zajmujące sią analizą danych.