Nr rejestracyjny: 2018/31/N/ST6/01209; Kierownik projektu: mgr Paweł Piasecki

Klasyfikacja szeregów czasowych z wykorzystaniem lasów podobieństwa i głębokich sieci neuronowych

(Streszczenie popularnonaukowe)

Paweł Piasecki

Obecnie mamy dostęp do coraz większej ilości danych różnych rodzajów. Na szczególną uwagę wśród nich zasługują szeregi czasowe. Za przykład tego typu danych mohą posłużyć: dzienna temperatura powietrza w Poznaniu lub rozmiar pewnego organizmu mierzony każdego miesiąca. Szeregi czasowe gromadzone mogą być podczas procesów biznesowych, biologicznych czy zabiegów medycznych. Na ogół zagadnienia związane z szeregami czasowymi oddziela się od pozostałych obszarów analizy danych, ponieważ obserwacje są uporządkowane w czasie i możemy z tego faktu czerpać dodatkowe informacje. W ciągu ostatnich 20 lat analiza szeregów czasowych rozwijana była bardzo intensywnie. Do najbardziej rozwiniętych jej obszarów zaliczyć możemy: klasyfikację, analizę skupień, indeksowanie, prognozowanie, detekcję anomalii, analizę wzorócw i inne.

W niniejszym projekcie skupimy się na obszarze klasyfikacji szeregów czasowych. Podamy w tym miejscu krótki przykład rozwijający intuicję na temat tego zagadnienia. Rozważmy zapis EKG pewnego pacjenta. Jest to szereg czasowy, ponieważ został zapisany w określonym przedziale czasu. Dane takie zazwyczaj wykorzystywane są w diagnozowaniu chorób serca. Zagadnienie związane z konstrukcją algorytmu wskazującego czy dana osoba posiada lub nie chorobą serca, jest zagadnieniem klasyfikacji szeregów czasowych.

W ciągu ostatnich dwóch dekad klasyfikacja szeregów czasowych postrzegana była jako jeden z najbardziej wymagających obszarów analizy danych. Wraz ze wzrostem dostępności danych, możliwe było rozwijanie kolejnych algorytmów. Obecnie istnieją już setki metod klasyfikacji. Jedną z najbardziej popularnych jest metoda najbliższych sąsiadów w połączeniu z różnymi miarami odległości. W szcególności często łączy się najbliższego sąsiada z odległością DTW (ang.*Dynamic Time Warping*). Wśród pozostałych klasyfikatorów na uwagę zasługują między innymi maszyna wektorów nośnych (ang. *Support Vector Machine*) oraz - popularne szczególnie w ostatnim czasie - sieci neuronowe.

Głównym zadaniem naszego projektu jest skonstruowanie nowego klasyfikatora dla szeregów czasowych - lasów podobieństwa. Podejście to zaproponowane zostało już dla danych wielowymiarowych i bazuje na lasach losowych. Lasy losowe w każdym węźle wybierają w sposób zrandomizowany cechy, na podstawie których przeprowadzany jest konstrukcja składowych drzew decyzyjnych. W przypadku szeregów czasowych nie dysponujemy cechami, które moglibyśmy losować. Musimy zatem takie cechy wygenerować. W tym celu lasy podobieństwa bazują na odległościach pomiędzy szeregami czasowymi.

Jak wynika z przeprowadzonych przez nas badań wstępnych, lasy podobieństwa wykazują większą skuteczność niż klasyfikator 1NN-DTW, określany w literaturze jako "trudny do pobicia". Co więcej, planujemy poprawić efektywność implementowanego algorytmu poprzez modyfikację metody wyboru cech w węzłach. W naszym projekcie chcemy porównać dokładność zaproponowanego przez nas finalnego klasyfikatora z istniejącymi już klasyfikatorami. Naszym celem jest także stworzenie zoptymalizowanej biblioteki R, tak aby zapewnić szeroki dostęp do rozwijanej metody naukowcom z różnych dziedzin.

W drugiej części naszego projektu chcemy przenieść podejście oparte na podobieństwach z lasów podobieństwa do sieci neuronowych. Uważamy, że podobieństwa pomiędzy szeregami czasowymi niosą ze sobą dużo informacji i należy to zagadnienie intensywnie zbadać. Początkowo planujemy użyć wektorów opartych na podobieństwach jako wejście do tradycyjnej sieci neuronowej. W kolejnym kroku rozwiniemy to podejście dla metod głębokiego uczenia i - jednej z najbardziej popularnych metod klasyfikacji szeregów czasowych w ostatnich latach - sieci LSTM (ang. *Long short-term memory network*).