

OPTIMALIZACJA, WALIDACJA ORAZ ZAKRES ZASTOSOWALNOŚCI JEDNOKLASOWYCH METOD KLASYFIKACJI

Zuzanna Małyjurek

Modelowanie jednoklasowe jest jednym z typów modelowania klasyfikacyjnego, stosowane w potwierdzaniu autentyczności, kontroli jakości i weryfikacji pochodzenia próbek. Modelowanie to służy do konstrukcji indywidualnych modeli klasowych bazując na podobieństwach między próbkami pochodzącymi z badanej klasy. Uzyskany model stosuje się do przewidywania czy nowa próbka o nieznanym pochodzeniu należy do modelowanej klasy. Modelowanie jednoklasowe to proces wieloetapowy, na który składają się m. in. konstrukcja zbioru treningowego i testowego, wybór metody modelowania, optymalizacja modelu oraz jego walidacja.

Celem badań omówionych w niniejszej rozprawie doktorskiej było wskazanie optymalnej strategii optymalizacji modelu klasowego, potwierdzania autentyczności klas podobnych do siebie oraz określenie zakresu stosowalności wybranych metod modelowania jednoklasowego.

W przypadku optymalizacji modelu klasowego, np. wyboru kompleksowości modelu, można wyróżnić dwie strategie. Jedna z nich bazuje jedynie na próbkach klasy modelowanej, a druga zakłada wykorzystanie również próbek spoza tej klasy. Wpływ tych dwóch strategii optymalizacji modelu na wyniki klasyfikacji oceniono na przykładzie modelowania różnych zbiorów danych metodą SIMCA (*ang. Soft Independent Modelling of Class Analogy*). Uzyskane wyniki wskazują, że modele optymalizowane z wykorzystaniem próbek modelowanej klasy oraz próbek spoza niej, prowadziły do najlepszych wyników klasyfikacji. Wykazano jednak, że modele te mogą prowadzić do zafałszowanych wyników, ponieważ reprezentatywność zbioru próbek spoza klasy modelowanej nie jest pewna. Zalecana jest zatem optymalizacja modelu klasowego bazując jedynie na próbkach modelowanej klasy.

Stosując optymalną metodę optymalizacji modelu, wybrane metody modelowania jednoklasowego zostały przetestowane względem ich zakresu stosowalności. Modele skonstruowane metodą SVDD (*ang. Support Vector Description Domain*) najczęściej prowadziły do najlepszych wyników klasyfikacji. Jednak w przypadku analizy danych o złożonej strukturze należy stosować metody oparte na gęstości próbek w przestrzeni parametrów np. metodę funkcji potencjalnych.

Indywidualne modele klasowe mogą prowadzić do niesatysfakcjonujących wyników, gdy weryfikowana jest autentyczność kilku podobnych do siebie klas. Wówczas do lepszych wyników klasyfikacji próbek prowadzą modele dyskryminacyjne, jednak one, w ich klasycznym ujęciu, nie mogą być stosowane do potwierdzania autentyczności próbek. W związku z tym, zaproponowano dwuetapową strategię potwierdzania autentyczności, która łączy modelowanie jednoklasowe i dyskryminacyjne. Skuteczność zaproponowanej strategii porównano z dwoma zmodyfikowanymi metodami dyskryminacyjnymi, które mogą być stosowane do weryfikacji autentyczności. Metody te, oprócz dyskryminacji próbek pomiędzy badane klasy, umożliwiają zidentyfikowanie próbki jako nienależącej do żadnej z badanych klas.

Wnioski uzyskane z powyżej omówionych badań nt. różnych etapów modelowania jednoklasowego, pozwoliły na zaproponowanie skutecznych strategii potwierdzania autentyczności herbaty z czerwonej kruszyny i miodokruszyny.