Nr rejestracyjny: 2019/35/B/ST6/04442; Kierownik projektu: prof. dr hab. inż. Michał Wojciech Woźniak

Wykorzystanie optymalizacji wielokryterialnej w uczeniu klasyfikatorów dla wybranych zadań decyzyjnych

Celem projektu jest zbadanie możliwości wykorzystania optymalizacji wielokryterialnej w zadaniach uczenia klasyfikatorów, w których kryterium uczenia obejmuje dwa lub więcej przeciwstawnych wskaźników. Przykładem tego typu zadania jest klasyfikacja danych niezbalansowanych, podczas której staramy się zbalansować *czułość* (metryka *recall*) klasyfikatora dla poszczególnych klas, w taki sposób, aby nie doprowadzić do nadmiernego pogorszenia jakości predykcji dla niedostatecznie reprezentowanych klas. Innymi słowy, w przypadku klasyfikacji danych niezbalansowanych mamy do czynienia z problemem, w którym koszt niepoprawnej klasyfikacji nie jest równy dla każdej z klas, a w większości przypadków nie jest też bezpośrednio zdefiniowany.

Standardem przyjętym w literaturze, dla dwuklasowych zadań klasyfikacji danych niezbalansowanych, jest ocena jakości predykcji przy pomocy metryk zagregowanych, takich jak AUC, F_β*-measure*, czy *G-mean*, obliczanych w oparciu o przeciwstawne kryteria typu *precision* i *recall*. Tego typu podejścia cechują się jednak kilkoma istotnymi wadami. W trakcie realizacji poprzednich projektów zaobserwowano, że wykorzystywanie wspomnianych kryteriów uczenia – w przypadku danych niezbalansowanych – prowadzi do utraty informacji o preferencjach modelu względem klas, ponieważ możliwe jest osiągnięcie tej samej wartości metryki zagregowanej dla wielu różnych kombinacji czułości osiąganych dla poszczególnych klas. Ponadto, optymalizując model względem metryk zagregowanych nie są uwzględniane preferencje użytkownika, ponieważ w przypadku optymalizacji jednokryterialnej, wybór konkretnego rozwiązania dokonywany jest w sposób arbitralny. Problem ten może zostać zniwelowany poprzez przypisanie kosztu niepoprawnej klasyfikacji dla każdej z klas, jednak w praktyce określenie tego kosztu *a priori* jest zwykle trudne, a poszukiwanie alternatywnych rozwiązań, jak *utility-based learning*, jest wciąż przedmiotem intensywnych badań.

Opisane problemy nie ograniczają się wyłącznie do klasyfikacji danych niezbalansowanych. Podobne obserwacje można poczynić dla zadania budowy zespołów klasyfikatorów, podczas którego staramy się dobrać klasyfikatory bazowe o wysokiej jakości predykcji, jak i dużej różnorodności, a także w klasyfikacji danych o zdefiniowanym koszcie akwizycji cech, podczas której konieczne jest zbalansowanie mocy predykcyjnej modelu oraz kosztu pozyskania konkretnych cech. Problem ten jest powszechny zwłaszcza w przypadku diagnostyki medycznej, gdzie do konstrukcji modelu decyzyjnego szukamy z jednej strony cech o dużej mocy dyskryminacyjnej, ale musimy również uwzględniać koszt pozyskania ich wartości, tj. odpowiedniego testu medycznego. Wielokryterialna natura widoczna jest też wreszcie w standardowych problemach uczenia klasyfikatorów, dla których powszechnie stosuje się metody regularyzacji, mające na celu zadbanie o zbalansowanie jakości predykcji na danych treningowych ze zbytnią złożonością modelu, która może prowadzić do przeuczenia (*overfitting*). W większości podejść, problem wielu kryteriów sprowadza się do konstrukcji zagregowanej funkcji celu uwzględniającej kryteria pojedyncze. Takie podejścia łączą wiele ograniczeń optymalizacji jednokryterialnej. Prowadzą do utraty informacji o zależnościach pomiędzy kryteriami składowymi, generują problem z uwzględnieniem preferencji użytkownika przy wyborze rozwiązania oraz rodzą trudności z interpretacją metryk zagregowanych.

W ramach projektu zbadana zostanie możliwość zniwelowania powyższych problemów przez wykorzystanie metod optymalizacji wielokryterialnej, zwracających zbiór Pareto-optymalnych rozwiązań, umożliwiających użytkownikowi wybór konkretnego modelu klasyfikacji oraz zaproponowane zostaną automatyczne metody jego wyboru, bądź agregacji modeli z wykorzystaniem paradygmatu klasyfikacji kombinowanej. W tym celu sformułowana została następująca hipoteza badawcza:

Możliwe jest opracowanie algorytmów uczenia klasyfikatorów wykorzystujących optymalizację wielokryterialną, zwracających zbiór Pareto-optymalnych klasyfikatorów, o indywidualnej jakości nie gorszej niż klasyfikatory wytrenowane przy użyciu optymalizacji jednokryterialnej.

W trakcie projektu opracowane zostaną metody uczenia klasyfikatorów wykorzystujących optymalizację wielokryterialną, zwracających zbiór reprezentatywnych, równoważnych w sensie Pareto rozwiązań, o jak największym stopniu zróżnicowania i jak najwyższej jakości. W jego ramach zostaną zaproponowane metody budowania pojedynczych modeli klasyfikacji oraz algorytmy konstrukcji i selekcji zespołów klasyfikatorów opartych o optymalizację wielokryterialną, które zostaną następnie przystosowane do zadania klasyfikacji niestacjonarnych strumieni danych oraz optymalizacji w trybie online. Opracowane metody zostaną ponadto wykorzystane w wybranych zadaniach decyzyjnych, takich jak klasyfikacja danych niezbalansowanych, budowa zespołów klasyfikatorów, przeciwdziałanie przeuczeniu, czy też selekcja i ekstrakcja atrybutów na potrzeby redukcji przestrzeni cech przeciwdziałającej zjawisku *klątwy wielowymiarowości*.