

Streszczenie pracy autorstwa mgr inż. Cezary Maszczyk zatytułowanej „Zastosowanie głębokiej eksploracji danych, w tym dyskretnego głębokiego uczenia, w indukcji reguł logicznych zawierających złożone warunki elementarne” napisane w języku polskim

Niniejsza praca dotyczy tematyki uczenia maszynowego, opartego na regułach oraz eksploracji danych, ze szczególnym uwzględnieniem indukcji złożonych zależności, wyrażanych w języku deskryptorów (warunków elementarnych). Głównym celem było odkrycie głębszych relacji, których standardowe algorytmy indukcji reguł, oparte na heurystykach, nie są w stanie wykryć. To z kolei może w pewnych sytuacjach przyczyniać się do poprawy interpretowalności uzyskiwanych modeli regułowych, umożliwiając uzyskanie bardziej zwięzłych i zrozumiałych opisów danych. Autor prezentuje w pracy trzy metody indukcji reguł, które umożliwiają odkrywanie złożonych zależności w danych. Każda z nich obsługuje aż trzy rodzaje danych: klasyfikacyjne, regresyjne i przeżyciowe. Opisane metody wpisują się w koncepcję głębokiego dyskretnego uczenia, opisaną w rozdziale 5. W przeciwieństwie do popularnych obecnie algorytmów subsymbolicznych, takich jak np. głębokie sieci neuronowe, te zaproponowane w pracy koncentrują się na tworzeniu zrozumiałych dla człowieka opisów danych w postaci zbiorów reguł decyzyjnych. Każda z trzech metod - **ComplexConditions**, **MofNRules** oraz **DeepRules** - prezentuje odmienne podejście do zagadnienia. Wszystkie jednak skupiają się na poszukiwaniu skomplikowanych zależności w danych, których zapis z użyciem prostych warunków (np. wiek > 30 lub kolor = czerwony), stosowanych przez tradycyjne algorytmy regułowe, byłby nadmiernie rozbudowany i mało czytelny. Cechą wspólną wszystkich opracowanych metod jest działanie w oparciu o popularną strategię sekwencyjnego pokrywania, gdzie każda nowa reguła indukowana jest na zbiorze niepokrytych dotąd przykładów. Minimalizuje to redundancję, pozwalając na tworzenie zwięzłych zbiorów reguł, gdzie każda z nich wnosi unikalną wartość.

Algorytm **ComplexConditions** stanowi odświeżone podejście do idei konstruktywnej indukcji sterowanej danymi. Bazując na dobrze opisanej w literaturze metodzie RuleKit [1] dodatkowo rozszerza ją o szeroki wachlarz warunków złożonych, takich jak warunki relacji atrybutów (np. wzrost > waga) lub wewnętrznych alternatyw (np. kolor $\in$ {czerwony, zielony, żółty}). Umożliwia to tworzenie bardziej zwięzłych opisów danych w porównaniu do powszechnie stosowanych algorytmów regułowych. W efekcie może to prowadzić do zredukowania złożoności wynikowego modelu, ułatwiając tym samym jego interpretację.

Druga z metod, **MofNRules**, stanowi rozwinięcie pierwszej, wprowadzając do niej nowy rodzaj warunków: M-z-N. Pozwalają one opisywać zależności w danych, do których opisu konieczne byłoby użycie wielu reguł z warunkami prostymi. Algorytm ten działa zgodnie z ideą konstruktywnej indukcji sterowanej hipotezami, generując reguły w procesie dwóch następujących po sobie przebiegów indukcji, rozszerzając przy tym pierwotną przestrzeń cech. Wyjątkową cechą metody **MofNRules** jest także to, że jako jedna z nielicznych obecnych

w literaturze, umożliwia tworzenie reguł zawierających w przesłankach, obok warunków M-z-N, także inne rodzaje warunków.

Trzecia i ostatnia z opisywanych metod, **DeepRules**, prezentuje odmienne spojrzenie na problem. Zamiast wyczerpująco przeglądać przestrzeń warunków złożonych, tak jak miało to miejsce w przypadku **ComplexConditions**, algorytm ten buduje rozbudowane wyrażenia logiczne w postaci CNF (koniunkcja alternatyw) i DNF (alternatywa koniunkcji). Wyrażenia te mogą być następnie upraszczane, zastępując ich poszczególne fragmenty równoważnymi logicznie warunkami złożonymi. Algorytm umożliwia w ten sposób uzyskiwanie reguł, zawierających wszystkie rodzaje deskryptorów stosowanych w poprzednich dwóch metodach (z pewnymi ograniczeniami), oferując przy tym mniejszą złożoność obliczeniową.

Kluczowym wnioskiem z przeprowadzonych eksperymentów jest stwierdzenie, że żadna z metod nie jest uniwersalnie najlepsza dla wszystkich typów problemów. Wyniki pokazują, że ich skuteczność jest w dużej mierze zależna od charakterystyki danych. W przypadku zbiorów, gdzie nie występują złożone wzorce, proponowane podejścia mogą czasem prowadzić do generowania nadmiernie skomplikowanych reguł. Przeprowadzone badania pokazały jednak, że opisane metody mają potencjał do generowania zbiorów reguł o mniejszej złożoności niż algorytmy referencyjne, a ich wyniki predykcyjne pozostają przy tym na ogół na zbliżonym poziomie. Świadczy to o tym, że są one w stanie opisać dane w sposób bardziej zwięzły, a przy tym równie dokładny. Na tej podstawie można podejrzewać, że są one zdolne do odkrywania pewnych głębszych zależności w danych, które są pomijane przez algorytmy referencyjne. Jest to dobrze widoczne w studiach przypadków problemów „MONK”. Wszystkie z nich są możliwe do opisanie z użyciem zbioru reguł w postaci DNF. Mimo to, algorytm RuleKit nie odnajduje w ich przypadku poprawnych reprezentacji klasy pozytywnej. Pokazuje to, że choć daną zależność można opisać za pomocą reguł z warunkami prostymi, nie oznacza to, że zostanie ona w praktyce odkryta przez algorytm. Dzieje się tak, ponieważ metody uczenia regułowego działają zwykle w oparciu o heurystyki, nie analizując pełnej przestrzeni rozwiązań, przez co niektóre z nich mogą zostać pominięte. Opracowane algorytmy, poprzez ukierunkowanie na poszukiwanie złożonych wzorców, zwiększają szansę na ich wykrycie, prowadząc często do bardziej zwięzłych i potencjalnie bardziej interpretowalnych opisów danych.

[1] Gudyś, A., Sikora, M., Wróbel, Ł. *RuleKit: A comprehensive suite for rule-based learning. Knowledge-Based Systems, 194:105480, 01 2020.*