Explainable Artificial Intelligence Based on Neuro-Fuzzy Approach in Application to Recommender Systems

Tomasz Rutkowski

Polish-Japanese Academy of Information Technology (PJAIT)

Supervisor: **dr hab. Jerzy Paweł Nowacki, prof. PJAIT**
Auxiliary Supervisor: **dr Radosław Nielek**

A thesis presented for the degree of
Doctor of Philosophy

POLISH-JAPANESE ACADEMY
OF INFORMATION TECHNOLOGY

Warsaw, 2020

Streszczenie

W ostatnich latach środowisko badaczy sztucznej inteligencji skupiało się przede wszystkim na zwiększaniu dokładności działania algorytmów. Modele oparte o głębokie sieci neuronowe wykazały spektakularnie dobre rezultaty, zwłaszcza w problemach rozpoznawaniu obrazów, automatycznego tłumaczenia tekstów czy rozpoznawania głosu. Użycie sieci typu GAN (Generative Adversarial Networks) umożliwiło z kolei automatyczne generowanie obrazów i muzyki. Dzięki rozwojowi tych technik, możemy obserwować ogromny postęp w diagnostyce medycznej czy autonomicznych samochodach. Problem w tym, że te modele działają na zasadzie „czarnych skrzynek". Oznacza to, że o ile działają one dokładnie, to bardzo trudno jest uzasadnić z czego dokładnie wynika predykcja lub rekomendacja. Istnieje jednak wiele przypadków, w których brak wyjaśnienia jest nieakceptowalny, ponieważ konsekwencje podjęcia błędnej decyzji są zbyt duże. Może to dotyczyć medycyny, prawa, finansów czy sterowania autonomicznymi samochodami. W przypadku relacji międzyludzkich człowiek wymaga od eksperta wyjaśnienia rekomendacji. Bardzo rzadko mamy do czynienia z takim poziomem zaufania, iż przyjmujemy rekomendacje, nawet ekspertów, bez wytłumaczenia dlaczego to jest najlepsza opcja ze wszystkich rozważanych i w jaki sposób ekspert uznał, że jest ona najlepsza. Wystarczy sobie wyobrazić przyjęcie rekomendacji o wycięciu organu, tylko dlatego, że ktoś stwierdził, że to jest najlepsze rozwiązanie. Skoro wymagamy wyjaśnień od innych ludzi to dlaczego nie wymagać ich od algorytmów?

Praca „Wyjaśnialna sztuczna inteligencja oparta o podejście neuronowo-rozmyte w zastosowaniu do systemów rekomendacji" koncentruje się na zaproponowaniu technik, które spowodują, że systemy rekomendacji mogą być zarówno dokładne, jak i wyjaśnialne.

Niniejsza rozprawa składa się z pięciu rozdziałów. Pierwszy rozdział opisuje wiele prób zdefiniowania czym jest wyjaśnialna sztuczna inteligencja (Explainable Artificial Intelligence) i przedstawia powiązane terminy, takie jak interpretowalność czy transparentność. Ponadto, rozdział pierwszy zawiera wprowadzenie do systemów rekomendacji oraz kwestii interpretowalności modeli uczenia maszynowego, w tym problem kompromisu między interpretowalnością a dokładnością. Opisane są również cel i tezy pracy:

- Cel: Zaproponowanie i zaimplementowanie wyjaśnialnych algorytmów w zastosowaniu do systemów rekomendyjnych
- Teza 1: Użycie podejścia neuronowo-rozmytego w systemach rekomendacyjnych opartych o informacje o obiekcie daje możliwość generowania zrozumiałych wyjaśnień dla każdej rekomendacji.

- Teza 2: Systemy rekomendacyjne oparte o podejście Neuronowo-Rozmyte mogą być jednocześnie dokładne, interpretowalne i transparentne.

W rozdziale 2 przedstawiono bardziej szczegółowo systemy neuronowo-rozmyte jako systemy rekomendacji, wraz z technikami uczenia i generowania reguł. Dzięki temu staje się jasne, że interpretowalność, jaką zapewniają systemy neuronowo-rozmyte może być właściwym punktem wyjścia do opracowania wyjaśnialnego systemu rekomendacyjnego. W dalszej części rozdziału wyróżniono dwie metody generowania reguł – Wang-Mendel i Nozaki-Ischibuchi-Tanaka. W rozdziale trzecim zaproponowano dwie nowe metody kodowania danych nominalnych oraz przedstawiono wiele wariantów systemów rekomendacji opartych o podejście neuronowo-rozmyte. Podstawowe trzy wersje systemu oznaczone są jako „Recommender A", „Recommender B" i „Recommender C".

Pierwszy z nich wykorzystuje tzw. metodę „Zero-Order Takagi-Sugeno-Kang" (ZO-TSK) i optymalizowany jest przez użycie algorytmu ewolucyjnego „Grey Wolf Optimizer". Zastosowane podejście pozwoliło na osiągnięcie wysokiej dokładności, zachowując rozsądnie niską liczbę interpretowalnych reguł. Drugi system, „Recommender B", pozwolił na rozszerzenie rozważań dla przypadków, w których wyjście systemu jest zbiorem rozmytym, a nie liczbą. Aby właściwie dobrać balans między liczbą reguł a dokładnością, zastosowano kryterium informacyjne Akaike, kryterium końcowego błędu predykcji oraz kryterium informacyjne Schwartza. W trzecim systemie, „Recommender C", użyto bardziej skomplikowanej T-Normy, zwanej Dombi, co skutkuje wprowadzeniem większej liczby parametrów. Dzięki temu system jest bardziej elastyczny, lecz jednocześnie bardziej złożony. Można argumentować, że „Recommender A" jest najprostszy i naturalnie interpretowalny, natomiast „Recommender C" jest dokładniejszy, ale w wyniku jego optymalizacji istnieje ryzyko utraty pewnego stopnia interpretowalności.

Eksperymenty związane z zaproponowanymi systemami potwierdzają istnienie kompromisu między interpretowalnością a dokładnością. Kryterium Akaike pozwala na dobranie optymalnej liczby interpretowalnych reguł, tak aby system był możliwie interpretowalny i dokładny jednocześnie. Wszystkie badania opisane w rozdziale trzecim zostały przeprowadzone w oparciu o zestaw danych MovieLens 10M, który zawiera informacje o ocenach filmów sporządzanych przez użytkowników. Jest to wiodący i uznawany przez środowisko zbiór danych do porównywania algorytmów rekomendacyjnych. Rezultaty wszystkich badań związanych z zastosowaniem podejścia neuronowo-rozmytego zawarte są w następujących publikacjach autora rozprawy:

- Rutkowski Tomasz et al.: On explainable fuzzy recommenders and their performance evaluation. International Journal of Applied Mathematics and Computer Science. Vol. 29. No. 3. pp. 595-610 (2019).
- Rutkowski Tomasz et al.: On explainable flexible fuzzy recommender and its performance evaluation using the Akaike information Criterion. ICONIP 2019. pp. 717-724 (2019).
- Rutkowski Tomasz et al: On explainable recommender system based on fuzzy rule generation techniques, In: Artificial Intelligence and Soft Computing, LNAI 11508, ICAISC 2019. Part I. Springer. pp. 358-372 (2019).
- Rutkowski Tomasz et al: A content-based recommendation system using neuro-fuzzy approach. 2018 International Conference on Fuzzy Systems (FUZZ-IEEE 2018). pp. 1-8 (2018).
- Rutkowski Tomasz et al: Towards interpretability of the movie recommender based on a neuro-fuzzy approach. In: Artificial Intelligence and Soft Computing, LNAI 10842, ICAISC 2018. Springer. pp. 752-762 (2018).

Czwarty rozdział opisuje założenia systemu rekomendacji, w którym do dyspozycji są przykłady historyczne tylko z jednej klasy (one-class classification). System oparto o rzeczywiste dane i postawiono problem, w którym wyjaśnialność jest krytyczna. O ile przy rekomendowaniu filmów konsekwencje złej rekomendacji są niskie – wystarczy zmienić film na inny, o tyle w inwestycjach giełdowych konsekwencje podjęcia złej decyzji mogą być bardzo kosztowne. Z tego powodu w systemach rekomendujących akcje na giełdzie nie wystarczą same rekomendacje, ale niezbędne jest również uargumentowanie dlaczego dana rekomendacja jest najlepsza oraz w jaki sposób model wygenerował taki wniosek. Do utworzenia wyjaśnialnego systemu rekomendacji użyto publicznych danych, publikowanych przez fundusze inwestycyjne o ich transakcjach przy użyciu kwartalnych zgłoszeń formularza 13F. Następnie wzbogacono je o dane podane przez firmy notowane na giełdzie. Dzięki temu dla każdej transakcji uzyskano 21 wartości odpowiadających kryteriom podejmowania decyzji inwestycyjnych. W oparciu o te dane statystycznie wyznaczono funkcje przynależności zbiorów rozmytych, które wykorzystywane są do opisu zrozumiałego dla użytkownika końcowego. Natomiast, żeby zapewnić dokładność oraz adaptacyjność systemu, dla każdej historycznej transakcji zaproponowano algorytm tworzący zbiór rozmyty w wielowymiarowej przestrzeni. Zamiast tworzyć model, który uogólnia proces podejmowania decyzji, próbując dopasować do niego jak najwięcej przykładowych danych, w przedstawionym w rozdziale czwartym, nowatorskim podejściu,

wszystkie dane tworzą model, a rozpoznawanie wzorców odbywa się na podstawie sąsiedztwa historycznych przykładów do rozważanego do rekomendacji nowego obiektu (w tym przypadku jest to akcja na giełdzie dostępna w danym czasie opisana 21 cechami).

To co wyróżnia tę metodę od dostępnych podejść, to wykorzystanie stopnia aktywacji reguły (oraz sposób jej wyznaczenia) jako miary sąsiedztwa, a tym samym podobieństwa obiektów. Taki system można przedstawić i uogólnić jako system neuronowo-rozmyty, aczkolwiek w przypadku, w którym do utworzenia modelu używa się przykładów tylko dla jednej klasy, nie ma możliwości zastosowania klasycznych metod douczania w oparciu o optymalizację miary błędu, czyli np. w oparciu o wsteczną propagację błędów), natomiast zakładając, że cechy opisujące obiekt, w kontekście systemów rekomendacyjnych, odpowiadają kryteriom podejmowania decyzji, można przyjąć, że należy rekomendować obiekty, które mają jak najwięcej, jak najbliższych przykładów wokół siebie. W rozdziale czwartym zaproponowano również metodę wizualizacji obiektów podobnych. Dzięki potraktowaniu stopnia aktywacji reguły jako miary odległości, możliwe jest przedstawienie na płaszczyźnie obiektów podobnych występujących w wielowymiarowej przestrzeni, bez konieczności redukcji wymiarów, a tym samym utraty wiedzy. Tę samą metodę można uogólnić dla przypadku wielu klas. Wówczas możliwe jest optymalizowanie systemu technikami przedstawionymi w rozdziale trzecim. Rozdziały 3 i 4 zawierają oryginalne wyniki badań oraz stanowią wkład autora do rozwoju systemów rekomendacji oraz metod wyjaśnialnej sztucznej inteligencji. Rozdział piąty zawiera podsumowanie i opis dalszych kierunków badań. Treść pracy jasno pokazuje, że osiągnięto cel badawczy oraz potwierdzono tezy.