



dr hab. Paweł Teisseyre, prof. IPI PAN

Instytut Podstaw Informatyki, Polska Akademia Nauk

teisseyrep@ipipan.waw.pl

Recenzja

rozprawy doktorskiej **mgr Kamila Kmity**

pt. „*Explainable Impact of Partial Supervision in Semi-Supervised Fuzzy Clustering*”

Promotor: **dr hab. Katarzyna Kaczmarek-Majer, prof. IBS PAN**

Tematyka pracy

Recenzowana praca dotyczy problemu częściowo nadzorowanej, rozmytej analizy skupień (Semi-Supervised Fuzzy Clustering, w skrócie: SSFC). Częściowe nadzorowanie oznacza, że dla pewnego podzbioru obserwacji dysponujemy jedynie wektorami cech, natomiast dla pozostałego zbioru mamy również informacje o przynależności do klas. Określenie „rozmyta” odnosi się do sytuacji w której dana obserwacja może należeć jednocześnie do kilku grup (skupień), ale z różnym stopniem przynależności. Na przykład, w zastosowaniach medycznych, pacjent może mieć objawy pasujące do kilku chorób naraz. Problematyka ta znajduje się na przecięciu takich obszarów jak: uczenie pod częściowym nadzorem (semi-supervised learning), analiza skupień (clustering) oraz logika rozmyta (fuzzy logic).

Doktorant nie proponuje nowych algorytmów rozmytej analizy skupień, skupia się natomiast na wpływie komponentu nadzorowanego na działanie istniejących algorytmów. Wpływ ten analizowany jest z wykorzystaniem narzędzi wyjaśnialnej sztucznej inteligencji (ang. Explainable Artificial Intelligence, XAI), co pozwala osadzić tematykę pracy w istotnym nurcie badań nad sztuczną inteligencją (AI) i uczeniem maszynowym (ML), mimo że sam problem częściowo nadzorowanej rozmytej analizy skupień dotyczy relatywnie wąskiego obszaru badawczego.

W pracy rozważa się algorytmy SSFC bazujące na funkcji celu, która jest sumą dwóch komponentów: składnika nienadzorowanego oraz składnika nadzorowanego (additive combination technique). Autor pokazuje, że w przypadku wielu istniejących metod SSFC (takich jak na przykład P97, GAN, SSFCM, SSPCM), funkcja celu może być zapisana właśnie w takiej formie. Komponent nienadzorowany odpowiada za identyfikację naturalnej struktury danych na podstawie podobieństwa między obiektami. Z kolei komponent nadzorowany wykorzystuje dostępną wiedzę a priori (etykiety obserwacji) dla części danych, aby ukierunkować proces grupowania i poprawić jakość uzyskanych skupień. Celem metod SSFC jest minimalizacja funkcji celu względem argumentów U (macierz przynależności) i V (macierz prototypów). Wpływ komponentu nadzorowanego jest kontrolowany za pomocą parametru α (scaling factor). Zwiększanie wartości α powoduje wzrost wpływu komponentu nadzorowanego na wartość funkcji celu, a zatem również na otrzymane rozwiązanie. Istniejące w literaturze i dokładnie opisane w pracy metody SSFC, różnią się sposobem definiowania składników funkcji celu. Autor skupia się na dwóch istniejących algorytmach: Semi-Supervised Fuzzy C-Means (SSFCM) oraz Semi-Supervised

Possibilistic C-Means (SSPCM). Pierwszy algorytm należy do grupy najczęściej wykorzystywanych rozwiązań typu *state of the art* w analizowanym obszarze badawczym.

W literaturze były prowadzone rozważania dotyczące parametru α oraz zaproponowano metody wyznaczania możliwych wartości tego parametru. Wyczerpujące podsumowanie znajduje się w Rozdziale 2.2.4. Autorzy większości prac ograniczają się do stwierdzenia, że parametr ten pozwala zrównoważyć komponenty funkcji celu. Brakuje jednak dokładniejszej analizy pokazującej jaki jest wpływ parametru α na rozwiązanie, tzn. na postać macierzy przynależności oraz prototypów dla popularnych metod SSFS. Brakuje również analiz pokazujących jakie są konsekwencje wyboru zbyt małej wartości lub zbyt dużej wartości α . W mojej opinii są to interesujące i ważne pytania badawcze. Dokładna analiza związku między wartością parametru α oraz otrzymanymi rozwiązaniami jest ważna dla zrozumienia działania poszczególnych metod.

Recenzowana rozprawa składa się z 6 rozdziałów, liczy 186 stron. Bibliografia składa się ze 170 pozycji. Rozdział pierwszy (Introduction) zawiera wprowadzenie do tematyki SSFC, motywację, cele badań oraz opis głównych osiągnięć pracy. Rozdział drugi zawiera przegląd metod rozmytej analizy skupień oraz częściowo nadzorowanej, rozmytej analizy skupień. W tej części opisano algorytmy Fuzzy C-means i Possibilistic C-Means jak również ich częściowo nadzorowane wersje: SSFCM oraz SSPCM. Rozdział trzeci zawiera najważniejsze wyniki pracy dotyczące wpływu parametru α na rozwiązania metod SSFCM oraz SSPCM. Rozdział czwarty został poświęcony efektywnej implementacji metod w języku R. Wyniki eksperymentów, opis sytuacji niedostatecznego oraz nadmiernego wpływu częściowego nadzoru, jak również sposoby wyznaczania siatki możliwych wartości parametru α zostały opisane w Rozdziale 5. Wnioski i podsumowanie znajdują się w Rozdziale 6.

Ocena merytoryczna pracy

Przegląd istniejących w literaturze metod SSFC, ich pogłębiona analiza oraz, co najważniejsze, analiza wpływu komponentu nadzorowanego na działanie poszczególnych metod pokazują szeroką wiedzę Autora w tej dziedzinie. Wyniki badań naukowych prowadzonych przez Doktoranta zostały opisane w 4 artykułach naukowych. Na szczególną uwagę zasługuje współautorska praca „Explainable Impact of Partial Supervision in Semi-Supervised Fuzzy Clustering”, opublikowana w renomowanym czasopiśmie IEEE TRANSACTIONS ON FUZZY SYSTEMS. Pan magister Kamil Kmita jest pierwszym autorem tej publikacji. Poniżej znajduje się lista najważniejszych w mojej ocenie osiągnięć opisanych w rozprawie.

1. Bardzo kompetentne i obszerne podsumowanie stanu wiedzy dotyczącej metod rozmytej analizy skupień oraz częściowo nadzorowanej analizy skupień. Dla każdej z metod, Autor analizuje funkcje celu, algorytm optymalizacji oraz postać rozwiązania.
2. Ważnym elementem pracy jest analiza zależności funkcyjnej między parametrem α a macierzą przynależności oraz macierzą prototypów (Rozdziały 3.3.2 oraz 3.3.3) dla algorytmu SSFCM i analogiczne wyniki dla algorytmu SSPCM. Rozważa się trzy przypadki: przynależność obserwacji nienadzorowanej do dowolnego skupienia, przynależność obserwacji etykietowanej do skupienia nienadzorowanego oraz przynależność obserwacji etykietowanej do skupienia nadzorowanego. W każdym z trzech wymienionych przypadków zależność funkcji przynależności od parametru α jest inna.
3. Interesującym i wartościowym elementem pracy jest również analiza zjawiska związanego z niedostatecznym wpływem częściowego nadzoru (underimpact of partial supervision). Polega ono na nieoczekiwanej zmianie kolejności wierszy w macierzach decyzyjnych V dla małych wartości parametru α . Zjawisko to jest związane z losową inicjalizacją algorytmu optymalizacyjnego. Niedostateczny wpływ częściowego nadzoru wpływa negatywnie na dokładność przewidywania etykiet. W pracy zaproponowano narzędzie diagnostyczne do detekcji takich sytuacji (wykres U-plot). Analizowano również sytuację nadmiernego wpływu częściowego nadzoru, polegającą na użyciu zbyt dużej wartości α . Opracowane narzędzie diagnostyczne (O-plot) pozwala zilustrować wpływ dużych wartości parametru α na

rozwiązanie. Opisane wyżej sytuacje zostały zilustrowane na przykładach rzeczywistych zbiorów danych. Przykłady te są właściwie dobrane i pozwalają zrozumieć istotę problemu. Wartościowa jest analiza częstości występowania efektu niedostatecznego wpływu częściowego nadzoru opisana w Rozdziale 5.4.2.

4. Analiza zjawisk niedostatecznego i nadmiernego wpływu komponentu nadzorowanego prowadzi do naturalnej propozycji ustalania siatki dopuszczalnych wartości parametru α , co ma duże znaczenie praktyczne. Zaproponowano metodę wyznaczania minimalnej oraz maksymalnej wartości parametru jak również liczbę możliwych wartości.
5. Na uwagę zasługuje również implementacja algorytmów w pakiecie fustclust, w języku R. Pakiet zawiera implementacje 4 metod: FCM, PCM, SSFCM, SSPCM. Efektywna implementacja nie jest zadaniem trywialnym; przyspieszenie obliczeń jest możliwe dzięki obliczaniu funkcji przynależności za pomocą operacji macierzowych a nie za pomocą formuł dla poszczególnych obserwacji.

Wszystkie powyższe uwagi skłaniają do bardzo pozytywnej oceny recenzowanej pracy.

Komentarze i pytania

1. W pracy zakłada się, że liczba klas określa liczbę rozpatrywanych skupień, tzn. każde skupienie odpowiada określonej klasie (strona 41: „It shall be noted that in this thesis we consider the classical, supervision approach, where the number of classes defines the number of clusters to be considered, i.e., each cluster corresponds to a specific class.”). To jest silne założenie. Łatwo sobie wyobrazić sytuację w której obserwujemy wiele wyraźnych skupień w obrębie jednej klasy. Istnienie skupień w obrębie jednej klasy uzasadniałoby przewagę metod analizy skupień nad użyciem metod klasyfikacji.
2. Czy zakłada się, że wszystkie klasy reprezentujące skupienia są obserwowane w zbiorze etykietowanym? Czy możliwa jest sytuacja w której nie mamy etykiet odpowiadających pewnemu skupieniu? Czy to będzie miało negatywny wpływ na działanie analizowanych metod?
3. Czy w problematyce SSFC rozpatruje się sytuacje w której dana obserwacja nadzorowana może należeć do kilku klas jednocześnie? To odpowiada sytuacji klasyfikacji wieloetykietowej (multi-label classification). Na przykład, u pacjenta rozpoznano kilka chorób jednocześnie. Czy rozpatrywane metody można zastosować w takiej sytuacji lub czy można je zmodyfikować aby obejmowały opisany przypadek?
4. W eksperymentach symulacyjnych (Rozdziały 1.4, Rozdział 5.1, Rysunek 1.3) zakłada się, że mechanizm etykietowania jest całkowicie losowy (strona 12: „We simulate partial supervision by retaining the class belonging information for only 20% of randomly sampled observations from each class”), co odpowiada rozważanemu w literaturze założeniu MCAR (Missing Completely at Random). Czy to założenie jest używane i czy jest potrzebne w rozważaniach teoretycznych? W literaturze rozważa się mniej restrykcyjne założenia, takie jak MAR (Missing at Random) czy MNAR (Missing Not At Random), gdzie prawdopodobieństwo etykietowania obserwacji może zależeć odpowiednio od wektora cech (MAR), lub zarówno od wektora cech jak i prawdziwej etykiety (MNAR). Wydaje mi się, że analiza takich schematów w eksperymentach symulacyjnych byłaby interesującym uzupełnieniem.
5. Zjawisko niedostatecznego wpływu komponentu nadzorowanego, dokładnie analizowane w Rozdziale 5, jest związane z niewłaściwym uporządkowaniem wierszy w wyznaczonej macierzy prototypów V , co przekłada się na niską dokładność klasyfikacji. W skrajnej sytuacji, dla $\alpha=0$, wykonujemy analizę skupień bez nadzoru, a więc przypisanie do klas może być wykonane z dokładnością do permutacji etykiet. Należy jednak podkreślić, że to zjawisko nie ma negatywnego wpływu na skuteczność wyznaczania skupień, mierzoną za pomocą miar takich jak Silhouette index (str 131: „The underimpact of partial supervision is not about quality of clustering, but about the unexpected reordering of rows in the decision matrices.”). To rodzi pytanie, na ile poważny jest to problem, w sytuacji kiedy celem analizy jest otrzymanie jak

najlepszych skupień a nie optymalizacja dokładności klasyfikacji (dla tego ostatniego zadania optymalnym podejściem jest użycie metod klasyfikacji pod częściowym nadzorem).

6. Bardzo doceniam, że w ramach przygotowania rozprawy doktorskiej został przygotowany pakiet R, zawierający implementacje opisanych metod. W ten sposób, zwiększamy prawdopodobieństwo tego, że metody zostaną użyte przez praktyków wykorzystujących częściowo nadzorowane metody analizy skupień. Czy Autor planuje przygotowanie implementacji również w języku Python, który jest obecnie najbardziej popularnym językiem używanym w środowisku ML?
7. W definicji 6 powinno być $p(y) := P(Y = y)$ bo nie jest jasne co oznacza $p(y)$ we wzorze (2.38).
8. Praca jest bardzo obszerna. Doceniam wnikliwą analizę różnych metod SSFC jednak pewne części pracy są moim zdaniem nadmiernie rozbudowane (np. Rozdziały 3.1.2, 4.1, 4.2), co nieco utrudnia skupienie się na najważniejszych fragmentach. Moim zdaniem, paragrafy dotyczące różnic między wyjaśnialnością i interpretowalnością (Rozdział 3.1.2) są zbyt obszerne i odbiegają od ścisłego, inżynierskiego sposobu prezentacji zagadnień. Rozdział dotyczący implementacji i pakietu R mógłby zostać przeniesiony do Dodatku.

Podsumowanie

Recenzowana praca dokumentuje szeroką wiedzę i kompetencje Doktoranta. Przedstawione w rozprawie analizy dotyczące wpływu komponentu nadzorowanego na działanie metod w sposób istotny uzupełniają aktualny stan wiedzy w obszarze częściowo nadzorowanej, rozmytej analizy skupień. Uwagi dyskusyjne i krytyczne nie pomniejszają jej wysokiej wartości. Stwierdzam, że rozprawa spełnia ustawowe warunki stawiane pracom doktorskim i wnioskuję o jej dopuszczenie do dalszych etapów przewodu doktorskiego.

Warszawa, 22/06/2026

Paweł Teisseyre

