Prof. dr hab. inż. Krzysztof Krawiec
Instytut Informatyki
Politechnika Poznańska
ul. Piotrowo 2
60-965 Poznań

Poznań, 21.05.2023

Recenzja rozprawy doktorskiej mgr Magdaleny Wiercioch p.t. “Development of universal data representations with application in chemistry”

1. Tematyka rozprawy

Przedmiotem rozprawy doktorskiej Pani mgr Magdaleny Wiercioch jest uczenie maszynowe w zastosowaniach w informatyce chemicznej. Głównym celem rozprawy jest projektowanie, implementacja i empiryczna ewaluacja modeli uczenia maszynowego zorientowanych na przetwarzanie informacji o strukturze związków chemicznych i realizacji zadań predykcji (klasyfikacji i regresji) wybranych właściwości fizykochemicznych tych związków. Rozważana klasa zadań ma bardzo istotne implikacje praktyczne i stanowi znaczne wyzwanie w informatyce chemicznej oraz interesujące pole badań koncepcyjnych dla uczenia maszynowego. W mojej ocenie praca jednoznacznie reprezentuje dyscyplinę Informatyka Techniczna i Telekomunikacja.

2. Treść rozprawy

Praca składa się z sześciu rozdziałów, załącznika oraz obszernej bibliografii. Po rozdziale 1, wprowadzającym w tematykę rozprawy i opisującym główne przyczynki pracy, w rozdziale 2 Autorka definiuje podstawowe pojęcia związane z tematyką rozprawy, szczególności zadanie uczenia się reprezentacji w problemach związanych z informatyką chemiczną. Kolejne rozdziały, od 3 do 5, prezentują trzy niezależne podejścia, a dokładniej modele złożonych głębokich sieci neuronowych, stanowiące główne przyczynki rozprawy, oraz wyniki demonstrujące skuteczność tych podejść w wybranych zastosowaniach praktycznych w informatyce chemicznej. Rozprawę zamyka podsumowanie, prezentujące także wskazania na możliwe kierunki dalszych prac.

Szczególny nacisk położony jest w rozprawie na uczenie się reprezentacji, co uważam za uzasadnione, ponieważ wiele zadań związanych z informatyką chemiczną wymaga od systemu uczącego się przyjęcia na wejściu złożonej reprezentacji związku chemicznego, o zmiennej długości i potencjalnie skomplikowanej strukturze. Stanowi to często znaczące wyzwanie i stanowi jedną z przyczyn dla których metody uczenia maszynowego zaczęły odnotowywać znaczące sukcesy w tym obszarze dopiero stosunkowo niedawno.

2.1. Główne przyczynki rozprawy

Głównymi przyczynami rozprawy są w mojej opinii trzy architektury głębokich sieci neuronowych zaprezentowane w rozdziałach 3, 4 i 5 rozprawy.

W rozdziale 3 Autorka proponuje architekturę o nazwie Hybrid Deep Neural Network (HybNN). Model ten przetwarza wejściową informację o strukturze cząsteczki związku chemicznego dwutorowo. Pierwszy z torów bazuje na grafowej sieci neuronowej wykorzystującej paradygmat propagacji komunikatów (*message passing*). Drugi tor przetwarza tę samą cząsteczkę chemiczną daną w postaci popularnej reprezentacji sekwencyjnej SMILES. Wektory zanurzeń zwracane przez obie te podsieci są następnie konkatenowane i podawane na wejście finalnej warstwy, która odwzorowuje pozyskane w ten sposób cechy na decyzję. Model wykorzystany został w tym rozdziale do predykcji aktywności związków chemicznych, do czego wykorzystano 14 benchmarków z repozytorium PubChem. Autorka skonfrontowała model HybNN z czterema metodami referencyjnymi, wykazując jego przewagę. Poza tym w rozdziale zilustrowano też przebieg procesu uczenia, zbadano czułość modelu na ustawienia wybranych hiperparametrów, oraz zweryfikowano hipotezę o związku pomiędzy oferowaną zdolnością predykcyjną a wzajemnym podobieństwem cząsteczek obecnych w benchmarku, mierzonym tak zwaną miarą podobieństwa Tanimoto. Wykazano też, za pomocą tzw. eksperymentu ablacyjnego, synergiczne role obu wymienionych wyżej torów. Wyniki eksperymentów wskazują na znaczną skuteczność i inne pozytywne charakterystyki podejścia HybNN.

Rozdział 4 prezentuje model nazwany przez Autorkę Subgraph Encoded Neural Network (SENN), służący do rozwiązywania zadań klasyfikacji lub regresji dla związków chemicznych. Punktem wyjścia jest założenie że tym razem graf reprezentujący molekułę jest grafem pełnym, gdzie atrybuty każdej krawędzi charakteryzują oddziaływanie między atomami. Kluczową cechą charakterystyczną proponowanego modelu jest bazowanie na fragmentach analizowanych grafów, które Autorka nazywa podgrafami k-ścieżkowymi (moje robocze tłumaczenie terminu *k-paths subgraphs*). SENN ekstrahuje z wejściowego grafu wszystkie takie podgrafy, a następnie odwzorowuje każdy z nich do uczalnej przestrzeni zanurzeń. Doktorantka przeprowadziła następnie empiryczną ocenę proponowanego modelu na 12 zadaniach klasyfikacji oraz 4 problemach regresji, porównując go z pięcioma metodami referencyjnymi w dla klasyfikacji oraz czterema dla regresji. Poza prezentacją i dyskusją tych wyników, zaprezentowane są także wyniki podobnych eksperymentów jak w rozdziale 3, a także analiza rozkładu błędu regresji. W końcowych sekcjach części eksperymentalnej rozdziału, Doktorantka zilustrowała pewne możliwości interpretacji proponowanego modelu przy pomocy znanej techniki atrybucji cech bazującej na scałkowanych gradientach (*integrated gradients*) oraz zaprezentowała wyniki eksperymentu

ablacyjnego polegającego na porównaniu z wariantem metody pozbawionym dostępu do cech fizykochemicznych. Wyniki eksperymentów przemawiają na korzyść architektury SENN i potwierdzają zasadność jego konstrukcji.

Ostatni przyczynkowy rozdział rozprawy, rozdział 5, opisuje autorską architekturę głębokiej sieci neuronowej Triplet Encoded Neural Network (TENN) przeznaczonej do predykcji interakcji pomiędzy lekami (*drugs*) a celami biologicznymi (*targets*). Proponowana architektura składa się z trzech podsieci, z których pierwsza przetwarza informację o formule substancji aktywnej (leku) reprezentowaną w postaci ciągu symboli SMILES, druga to sieć neuronowa analizująca cząsteczkę białka (target) reprezentowaną w postaci sekwencji tokenów (w sposób analogiczny do sekwencji w języku naturalnym), a trzecia to podsieć odpowiedzialna za zanurzenie reprezentacji leku reprezentowanego jako graf. Reprezentacje wypracowane przez te trzy komponenty są następnie konkatenowane oraz podawane na wejście liniowego agregatora, który z kolei produkuje finalną decyzję odnośnie obecności interakcji. Architektura TENN buduje po części na komponentach zaproponowanych przez Doktorantkę w poprzednich rozdziałach, w szczególności na komponentach do przetwarzania podgrafów k-ścieżkowych. Proponowane podejście zostało przetestowane na pięciu problemach pozyskanych z publicznej bazy danych BindingDB. Wyniki scharakteryzowane zostały w kategoriach znanych metryk (trafność klasyfikowania, precyzja, czułość, miara F1, pole powierzchni pod krzywą ROC). Podobnie jak w poprzednich rozdziałach, wyniki eksperymentów przemawiają na korzyść proponowanego algorytmu. Autorka przeprowadziła także eksperymenty strojenia wybranych hiperparametrów modelu, w szczególności wymiarowości zanurzenia oraz intensywności stosowania operacji dropout w trakcie uczenia.

2.2. Ocena treści rozprawy

Organizacja rozprawy jest dobrze przemyślana, a jej cele klarownie wyznaczone. Praca jest też tematycznie zwarta i w większości prezentuje treści na odpowiednim poziomie szczegółowości. Głównym atutem rozprawy jest wysoka jakość osiągniętych wyników empirycznych. Modele proponowane przez Doktorantkę osiągają lepsze wartości metryk na większości z rozważanych w pracy problemów klasyfikacji i regresji charakterystycznych dla informatyki chemicznej, co zawdzięczają moim zdaniem głównie (i) zaangażowaniu wyrafinowanych modeli neuronowych, zdolnych do przetwarzania złożonych struktur danych charakterystychnych dla chemii, oraz (ii) korzystaniu z dopełniających się źródeł informacji i alternatywnych reprezentacji (SMILES, grafy, wektory cech fizykochemicznych, sekwencje tokenów). Wskazuje to na zasadność wyborów podjętych przez Autorkę w procesie projektowania tych modeli oraz dobrą znajomość obranego obszaru badań. To ostatnie sugerują także rzeczowe przeglądy literatury prezentowane w poszczególnych rozdziałach pracy. **W związku z powyższym moja ogólna ocena pracy jest zdecydowanie pozytywna.**

Podczas lektury pracy dopatrzyłem się kilku pomniejszych niedociągnięć i niejasności, które omawiam poniżej.

Punkt 4 listy na stronie 23 sugeruje że ręcznie projektowane cechy pozbawione są zdolności i generalizacyjnej i skalowalności, podczas gdy w wielu przypadkach to właśnie takie cechy są nieodzowne dla skutecznego działania systemu uczącego się, na przykład gdy implementują wiedzą dziedzinową nieobecną jawnie w danych uczących. Być może Autorce chodzi tutaj raczej o, istotnie realne w przypadku ręcznego projektowania cech, ryzyko obciążania (*bias*) cech (i w konsekwencji modeli), czy to przez ich wybór (spośród potencjalnie szerszego spektrum cech), czy też przez sposób ich reprezentowania. Zauważmy jednak że źródłem obciążeń mogą być także dane uczące.

W sekcji 3.3.1, na stronach 37-38, Autorka deklaruje wykorzystywanie kodowania "gorącej jedynki" (*one-hot*) jedynie dla zmiennej reprezentującej typ atomu. Jednak z tabeli 3.1 widać że przynajmniej jedna inna cecha, *hybridization*, wydaje się być zdefiniowana jest na skali nominalnej (kategorycznej). Czy ona także kodowana jest w taki sposób? Natomiast kodowanie/reprezentacja pozostałych cech scharakteryzowane jest nie do końca jednoznacznym zdaniem "In all other cases, the value is assigned to its fixed index position". Domyślam się że Autorka sugeruje w ten sposób że wielkości te są podawane są na osobne, dedykowane dla nich pojedyncze wejścia modelu. Dla kompletności byłoby także wskazane wspomnieć czy/jak normalizowane są te cechy (domyślam się że z racji fizykochemicznego charakteru mają one zapewne bardzo różne zakresy).

Wyniki eksperymentalne prezentowane w sekcji 3.4 dotyczą 14 problemów klasyfikacji binarnej pochodzących z bazy danych PubChem. Baza ta zawiera obecnie około pół miliona problemów. Nawet jeśli jedynie część z nich jest problemami klasyfikacji binarnej, nasuwa się oczywiste pytanie: jakimi przesłankami Doktorantka kierowała się w wyborze tych właśnie konkretnych problemów? Wydaje się że motywacje te nie zostały przedstawione w tekście rozprawy.

W komponencie 1 (Component 1) podejścia TENN, opisywanym w sekcji 5.3.1, pojawiają się trzy następujące po sobie warstwy liniowe (Rys. 5.2). Autorka użyła wcześniej podobnej konstrukcji w architekturze SENN (Rys. 4.1), tj. dwóch warstw liniowych w końcowych stadiach modelu; tam jednak połączone było to z wykorzystaniem operacji dropout jedynie w pierwszej z tych warstw, co uzasadnia taką konstrukcję. Natomiast w TENN jedynym dodatkowym komponentem 'przeplatanym' z warstwami liniowymi wydaje się być normalizacja wsadowa (paczkami; *batch normalization*) – ani rysunek, ani towarzyszący mu opis nie wydaje się wskazywać na wykorzystanie operacji dropout, czy też innych komponentów (np. nieliniowych funkcji aktywacji). W konsekwencji użycie złożenia (kompozycji) warstw liniowych jest dyskusyjne, bo mogłyby być ono zastąpione jedną warstwą liniową (normalizowanie sygnałów nie podważa moim zdaniem zasadności niniejszej obserwacji).

Załącznik A w mojej ocenie jest nieco dyskusyjny, ponieważ prezentuje podstawowe pojęcia uczenia głębokiego, dostępne w wielu podręcznikach i innych źródłach. Niemniej niewątpliwie czyni on tę rozprawę bardziej niezależną (w sensie *self-contained*).

Wymienione tu uwagi są jak widać nieliczne i mają częściowo polemiczny charakter, a zatem nie podważają mojej ogólnie dobrej oceny pracy. Dodam też że moje wcześniejsze uwagi redakcyjne zostały one uwzględnione w ostatecznej wersji pracy.

3. Konkluzja końcowa

Rozprawa doktorska mgr Magdaleny Wiercioch zawiera wartościowe przyczynki do uczenia się reprezentacji w uczeniu maszynowym i zastosowań tych technik w informatyce chemicznej. Głównymi atutami rozprawy są (i) oryginalne modele uczenia maszynowego, korzystające z uzupełniających się źródeł informacji i różnych reprezentacji (SMILES, grafy, wektory cech fizykochemicznych, sekwencje tokenów), (ii) wysoka jakość uzyskanych wyników empirycznych, oraz (iii) wykazanie skuteczności proponowanych modeli na relatywnie zróżnicowanej gamie zastosowań.

Wobec powyższego stwierdzam, że **rozprawa doktorska mgr Magdaleny Wiercioch zdecydowanie spełnia warunki stawiane przez Ustawę Prawo o Szkolnictwie Wyższym i Nauce (Dz. U. z 2020 r., poz 85 z późn. zm.) w odniesieniu do rozpraw doktorskich, a zatem powinna być dopuszczona do publicznej obrony, o co wnoszę do Rady Dyscypliny Naukowej Informatyka Techniczna i Telekomunikacja Uniwersytetu Jagiellońskiego.**