Streszczenie

Methods for selected problems in unsupervised graph representation learning (Metody dla wybranych problemów nienadzorowanego uczenia reprezentacji grafów)

Metody uczenia maszynowego (ML) były badane w różnych zastosowaniach dla różnych rodzajów danych. Ponieważ większość modeli uczenia maszynowego oczekuje na wejściu wektora z ciągłej przestrzeni, zaproponowano metody uczenia reprezentacji, które automatycznie tworzą wektory reprezentacji (osadzenia) dla danych wejściowych. Podczas gdy istnieje wiele metod osadzania dla tradycyjnych typów danych, takich jak BERT dla tekstu czy ResNet dla obrazów, zadanie to jest znacznie trudniejsze dla danych o strukturze grafu. Grafy mogą być używane do opisania obiektów (za pomocą węzłów) i ich relacji (za pomocą krawędzi). Dla prostych grafów głównym celem metod uczenia reprezentacji jest uchwycenie struktury grafu. Grafy jednakże mogą zawierać dodatkowe informacje, tzn. oprócz struktury (węzłów i krawędzi), można przypisać atrybuty do węzłów, krawędzi, a nawet całych grafów (grafy atrybutowane). Budowa odpowiednich wektorów reprezentacji dla grafów jest kluczowa dla osiągnięcia dobrej jakości w zadaniach uczenia maszynowego, np. klasyfikacji węzłów czy predykcja powiązań.

W ostatnich latach zaproponowano różne metody uczenia reprezentacji grafów. Pomimo sukcesu tych metod, nadal nie rozwiązano wielu problemów i nie zapełniono luk badawczych. Przede wszystkim większość proponowanych podejść jest transduktywna (optymalizują macierz osadzeń o stałym rozmiarze), co nie pozwala na uzyskanie osadzeń dla wcześniej niewidzianych przykładów. Co więcej, dla rzeczywistych, dużych grafów z milionami, a nawet miliardami węzłów, to rozwiązanie wymaga ogromnej ilości pamięci i jest niepraktyczne, ponieważ grafy mogą ewoluować w czasie. Kolejnym poważnym wyzwaniem jest to, że ustawienie semi-nadzorowane wymaga danych oznaczonych, których pozyskanie jest drogim i czasochłonnym zadaniem. Większość istniejących podejść uczenia reprezentacji grafów zostało wprowadzonych w tym właśnie ustawieniu. Model jest optymalizowany pod konkretne zadanie docelowe, co ogranicza ekspresywność reprezentacji. Wykorzystanie alternatywnego podejścia uczenia nienadzorowanego i samo-nadzorowanego pozwoliło uzyskać reprezentacje grafów wyłącznie ze struktury sieci i atrybutów. Kolejnym wyzwaniem jest definicja negatywnych próbek w kontrastowych funkcjach kosztu używanych w metodach uczenia reprezentacji grafów. Jakość osadzeń, a co za tym idzie, wydajność w zadaniach docelowych, w dużej mierze zależy od tego, jak zdefiniowane są próbki negatywne. W porównaniu do innych typów danych wybór odpowiednich próbek negatywnych jest szczególnie trudny w grafach. Podjęto dlatego trud badawczy skupiony na metodach nie wykorzystujących próbek negatywnych (*negative-sample-free*), np. w przetwarzaniu obrazów i przetwarzaniu języka naturalnego. Jednakże grafy nie były przedmiotem takiego rozpoznania.

Rozprawa doktorska zakładała zrealizowanie następujących celów: (1) sprawdzenie, czy wykorzystanie miary korelacji krzyżowej zastosowanej do zmodyfikowanych widoków grafu atrybutowanego pozwala na trenowanie grafowej sieci neuronowej w sposób samo-nadzorowany, tak aby sieć dostarczała lepsze wektory reprezentacji węzłów niż inne istniejące podejścia samo-nadzorowane (pod względem jakości w zadaniach docelowych i złożoności czasowej procesu uczenia), 2) opracowanie modelu głębokiej sieci neuronowej do wyznaczania wektorów reprezentacji krawędzi, który jest trenowany przy użyciu połączenia kontrastowej funkcji kosztu z funkcją rekonstrukcji cech, tak aby wektory osadzenia tej metody były lepsze w zadaniach docelowych niż te uzyskane jako agregacje reprezentacji węzła źródłowego i docelowego, 3) opracowanie przyrostowej metody uczenia reprezentacji dla węzłów w grafach dynamicznych, która wykorzystuje dowolne statyczne metody osadzania węzłów z kolejnych migawek grafu i wykazuje niższą złożoność czasową i pamięciową niż istniejące metody uczenia reprezentacji na grafach dynamicznych, zapewniając jednocześnie przyrosty miary jakości, oraz 4) opracowanie metody łączenia strukturalnych osadzeń węzłów z informacjami o ich atrybutach w celu otrzymania pojedynczego niskowymiarowego osadzenia, które działa lepiej w zadaniach docelowych niż: strukturalne osadzenia, atrybuty węzłów czy inne metody uczenia reprezentacji

węzłów.

Rozprawa doktorska jest zbiorem powiązanych tematycznie prac w postaci pięciu publikacji naukowych skupionych wokół tematu nienadzorowanych metod uczenia reprezentacji grafów.

Pierwsza publikacja (*Graph Barlow Twins: A self-supervised representation learning framework for graphs*) proponuje nowy framework do samo-nadzorowanego uczenia reprezentacji węzłów w atrybutowanych grafach, który wykorzystuje empiryczną macierz korelacji krzyżowej pomiędzy osadzeniami dwóch zmodyfikowanych widoków grafu. Zapewnia to prostą i ekspresywną symetryczną architekturę sieci neuronowej, która nie wymaga negatywnych próbek w procesie uczenia. Eksperymentalna ewaluacja uwidoczniła, że otrzymywane reprezentacje węzłów osiągały analogiczną jakość w porównaniu do najnowszych metod, wymagając znacznie mniejszej liczby hiperparametrów i zbiegając szybciej, co prowadzi do ogólnego przyspieszenia do 42 razy w porównaniu do najnowszych metod.

W drugim artykule (*AttrE2vec: Unsupervised attributed edge representation learning*) wprowadzono nową metodę uczenia wektorów reprezentacji krawędzi. Metoda ta bada sąsiedztwo krawędzi za pomocą spaceru losowego i stosuje funkcję agregacji do uzyskania podsumowań sąsiedztwa, które następnie są przekazywane razem z atrybutami krawędzi do modułu kodera głębokiej sieci neuronowej. Model jest optymalizowany za pomocą złożonej funkcji straty składającej się z kontrastowego uczenia (aby uchwycić strukturę grafu) i funkcji rekonstrukcji cech (aby zapewnić, że atrybuty krawędzi są zakodowane prawidłowo w wektorze reprezentacji). Eksperymenty wykazały, że proponowana metoda tworzy reprezentacje, które osiągnęły lepszą jakość w zadaniach klasyfikacji i grupowania krawędzi w porównaniu do innych podejść najnowszej generacji.

Trzecia publikacja (*FILDNE: A Framework for Incremental Learning of Dynamic Networks Embeddings*) proponuje nowy framework do uczenia wektorów reprezentacji węzłów w dynamicznych grafach za pomocą uczenia przyrostowego. Model wykorzystuje dowolną podaną statyczną metodę uczenia reprezentacji węzłów i stosuje ją do dynamicznego grafu (modelowanego jako sekwencja migawek grafu). Następnie framework agreguje wektory osadzeń z kolejnych migawek za pomocą funkcji liniowo-wypukłej, której parametry są wyznaczane za pomocą modelu Dirichlet-Multinomial przy użyciu nienadzorowanego zadania predykcji połączeń. Eksperymentalna ewaluacja wykazała, że proponowany framework jest bardziej wydajny pod względem złożoności czasowej i pamięciowej, a jednocześnie osiąga lepszą jakość w porównaniu do współczesnych metod uczenia reprezentacji na dynamicznych grafach.

Czwarty artykuł (*Retrofitting Structural Graph Embeddings with Node Attribute Information*) oraz piąty (*A deeper look at Graph Embedding RetroFitting*) zajęły się problemem aktualizacji (retrofitowania) wstępnie obliczonych strukturalnych wektorów reprezentacji węzłów za pomocą informacji o ich atrybutach. Najpierw wprowadzono metodę opartą na złożonej funkcji celu, składającą się ze straty niezmienności (aby zachować informacje o osadzeniu strukturalnym), straty sąsiedztwa grafu (aby zwiększyć podobieństwo wektorów węzłów połączonych krawędziami) i straty sąsiedztwa atrybutów (aby zwiększyć podobieństwo wektorów węzłów o podobnych atrybutach). Ewaluacja eksperymentalna wykazała, że proponowana metoda osiągnęła lepsze wyniki w zadaniu klasyfikacji węzłów niż inne metody uczenia reprezentacji węzłów z atrybutami. Jednakże zaproponowana metoda posiadała wiele hiperparametrów, a funkcja celu była nadmiarowa. W związku z tym w piątym artykule zaproponowano uproszenie metody w zakresie funkcji celu i zaproponowano algorytm do automatycznego wyznaczania hiperparametrów. Rozszerzona ewaluacja eksperymentalna wykazała, że nowa metoda obliczała lepsze wektory reprezentacji niż istniejące podejścia do grafów z atrybutami.

Podsumowując, wyniki badań przeprowadzonych w ramach rozprawy doktorskiej wykazały, że zaproponowane metody i podejścia, jako nienadzorowane metody uczenia reprezentacji, dostarczają bardziej generalizujące wektory reprezentacji podczas ewaluacji na zadaniach docelowych niż dotychczas istniejące metody nienadzorowane.