Neighbor Embedding in Feature Selection and Multipoint Extensions

Adrian Łańcucki

21 września 2018

Streszczenie

Probabilistyczne metody redukcji wymiarowości to szczególne metody nowoczesnego przetwarzania danych. Stochastyczne zanurzenia sąsiedztw (*Stochastic Neighbor Embedding, SNE*) oraz stochastyczne zanurzenia sąsiedztw z rozkładem t-Studenta (*t-Distributed SNE*), to przykłady tychże cieszące się popularnością zarówno w środowiskach naukowych, jak też pośród osób zajmujących się zawodowo przetwarzaniem danych. Badania nad tymi metodami przeprowadzone w ciągu ostatnich kilku lat skupiały się głównie na szybkich, aproksymowanych algorytmach wyliczania zanurzeń oraz na redukowaniu wymiarowości jednego zbioru danych do postaci kilku zanurzeń (map).

Rozprawa dotyczy problemu kopiowania wybranych punktów konstruowanych zanurzeń w metodach zanurzeń sąsiedztw. Istniejące algorytmy, takie jak wspomniane zanurzanie w postaci kilku map (*multiple-map t-SNE*), zanurzanie jednego punktu w postaci kilku niskowymiarowych punktów (*mixture SNE*) oraz inne, zakładają stałą liczbę niskowymiarowych zanurzeń o wagach sumujących się do 1. Takie założenie upraszcza rachunki i obliczenia macierzowe, nie jest jednak realistyczne, biorąc pod uwagę rozkład Pareto (prawo potęgowe, *power law*). Rozkład ten modeluje sytuacje, w których nieliczna mniejszość kontroluje większość pewnego zasobu. W wypadku redukcji wymiarowości, za zasób można uważać liczbę punktów leżących w bliskim sąsiedztwie rozpatrywanego. Rozkład Pareto występuje tam, gdzie dochodzi do zjawisk o charakterze kolektywnym. Przykładowo, można nim modelować częstość występowania słów w językach naturalnych, ilość powiązań w sieciach społecznościowych, bogactwa zgromadzonego przez jednostki i wszystkie inne sytuacje, w których przejawia się występowanie tzw. zasady Pareto, w myśl której za 80% osiąganego efektu odpowiedzialne jest 20% włożonego wysiłku. Dane posiadające taki rozkład są trudne do wizualizacji przy użyciu istniejących metod redukcji wymiarowości. Świadome kopiowanie jedynie wybranych punktów, dokładnie tych, które skupiają wiele zasobów, mogłoby być pomocne w wizualizacji takich danych. Dodatkowo, charakter niektórych rodzajów danych wymaga wręcz kopiowania jedynie wybranych punktów, np. zanurzenia polisemicznych słów w językach naturalnych.

Wagowanie punktów w istniejących metodach redukcji wymiarowości jest uciążliwe w interpretacji, szczególnie dla dużych zbiorów danych, dla których ocenie nie podlegają już pojedyncze punkty, a ich skupiska. Nie jest bowiem jasne, jakie znaczenie powinno być nadane ułamkowi punktu, lub fragmentarycznym skupiskom. Jednym z celów niniejszej

rozprawy jest zaproponowanie i zbadanie rozszerzenia metod redukcji wymiarowości poprzez zanurzenia sąsiedztw, która tworzy pełnowartościowe kopie punktów o identycznej wadze, jak ich pierwowzory. Dodatkowo, kopiowaniu podlegają tylko wybrane punkty, a liczba kopii waha się pomiędzy różnymi punktami. Kopie są tworzone on-line podczas gradientowej optymalizacji funkcji kosztu, a decyzja o kopiowaniu zapada w oparciu o te wyliczone już gradienty. W ten sposób optymalizacja gradientowa, uważana dotąd za dużą wadę zanurzeń sąsiedztw, wykorzystywana jest na korzyść konstruowanego zanurzenia.

Drugorzędnym celem pracy jest demonstracja możliwości wykorzystania zanurzeń sąsiedztw poza wizualizacją danych, w konstruowaniu lepszych reprezentacji osobników w ewolucyjnych, spowitych metodach selekcji atrybutów (*wrapper feature selection*). Brakuje jednoznacznych wytycznych co do enkodowania potencjalnych rozwiązań w formie wektorów osobników, a istniejące praktyki opierają się często na pojedynczych przykładach opisanych w literaturze. W rozprawie rozważone zostały dwa przykłady wykorzystania zanurzeń sąsiedztw w spowitej selekcji atrybutów, oraz wyniki ich zastosowania w problemach obliczeniowych z dziedziny biologii molekularnej.

Główny wkład rozprawy w rozwój redukcji wymiarowości to algorytm wielopunktowego zanurzenia t-SNE. W prezentowanych testach obliczeniowych wykonanych na różnorakich danych (tj. danych dotyczących sieci społecznościowych, zdjęć, czy zanurzeń słów języka naturalnego) osiąga znacznie lepsze rezultaty, niż czołowe metody wizualizacji. Niskowymiarowe zanurzenia zachowują znacznie więcej pierwotnej informacji, mierzonej jakością rekonstrukcji lokalnych sąsiedztw punktów. Dwuwymiarowe wizualizacje przedstawione w rozprawie wyglądają istotnie lepiej, biorąc pod uwagę liczbę występujących artefaktów, niż te same dane zwizualizowane przy pomocy zwykłego algorytmu t-SNE. Opisane metody zanurzeń sąsiedztw są umieszczone w szerszym kontekście nowoczesnych metod uczenia maszynowego i stanowią nowy, interesujący kierunek badań. Dodatkowo została zaproponowana adaptacja algorytm szybkiej aproksymacji gradientów używanego w t-SNE na potrzeby wielopunktowego t-SNE. Zastosowania obu tych algorytmów w ewolucyjnej spowitej selekcji atrybutów pokazują ich potencjał w uproszczeniu reprezentacji dla tych metod, przy wykorzystaniu niskowymiarowych map atrybutów.