Recenzja pracy doktorskiej pana Adriana Łańcuckiego

Neighbor Embedding in Feature Selection and Multipoint Extensions

Rozprawa dotyczy różnych wariantów metody zanurzeń sąsiedztw, ze szczególnym uwzględnieniem zagadnienia kopiowania punktów w niskowymiarowej reprezentacji. Metoda zanurzeń sąsiedztw została zaproponowana w pracy Hintona i Roweisa *Stochastic Neighbor Embedding* (SNE) i stanowi przykład algorytmu pozwalającego na odwzorowanie zbioru danych do niskiego wymiaru, najczęściej do wymiaru 2 lub 3. .

W rozdziale 2 recenzowanej rozprawy jest przedstawiony szeroki przegląd metod związanych z redukcją wymiarowości. Przegląd jest wyczerpujący: podany jest wzór na włożenie lub zarysowany jest algorytm dający włożenie i następnie podsumowane są podstawowe empiryczne własności danego włożenia, a w niektórych przypadkach także teoretyczne własności. Włożenia w przestrzeń dwuwymiarową są dobrze zaprezentowane w tym rozdziale przy pomocy licznych ilustracji. Warianty metody zanurzeń sąsiedztw są starannie przedstawione w podrozdziale 2.2.3.

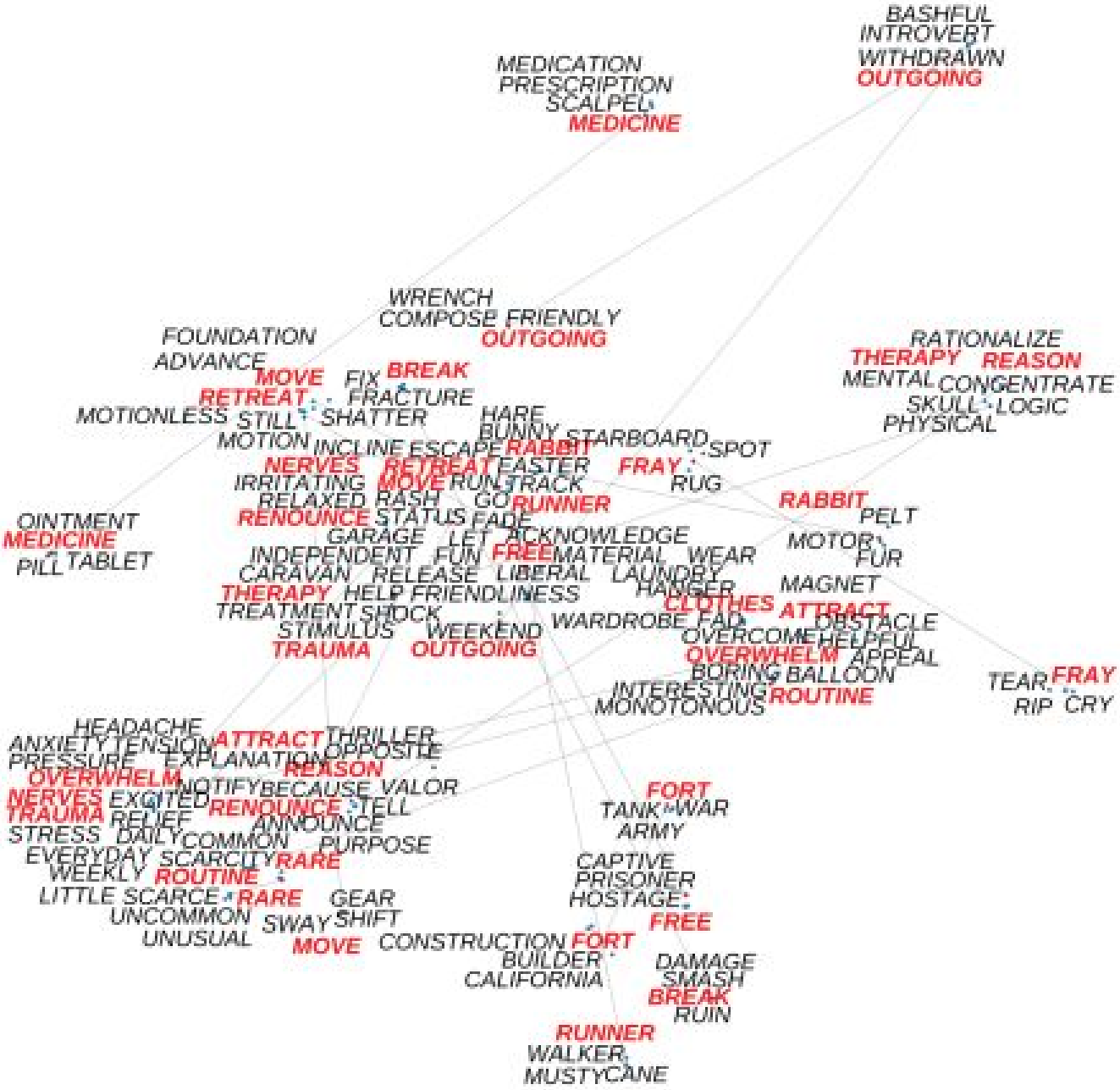
Algorytm wielopunktowego zanurzenia t-SNE

W rozdziale 3 recenzowanej rozprawy przedstawiony jest algorytm z pracy A. Łańcuckiego i J. Chorowskiego *Multipoint Neighbor Embedding* zaprezentowanej na konferencji International Conference on Text, Speech, and Dialogue w roku 2017. Wielopunktowe zanurzenia polegają na skopiowaniu niektórych punktów przy włożeniu. Aby zmierzyć jakość włożenia autorzy obliczają średnią precyzję z jaką k sąsiednich punktów jest zachowanych przy włożeniu.

Algorytm iteracyjnie dodaje nowe punkty tam gdzie funkcja kosztu wskazuje, że są najbardziej potrzebne i następnie usuwa te punkty, które są zbędne. Algorytm w swej podstawowej formie ma złożoność kwadratową względem liczby punktów we wkładanym zbiorze. W rozdziale 3.2.1 podana odpowiednio dostosowana metoda Barnesa-Huta przybliżenia gradientu funkcji celu. Dzięki temu sposobowi cały algorytm osiąga złożoność O(N log N), gdzie N to liczba punktów we wkładanym zbiorze.

W rozdziale 3.3 jest przedstawione porównanie włożenia wielopunktowego t-SNE z innymi, znanymi z literatury wariantami metody t-SNE względem średniej precyzji włożenia dla $k < 15$. Dla wszystkich analizowanych zbiorów danych włożenia wielopunktowe daje lepsze efekty, niż wcześniej znane metody. W tym samym rozdziale przedstawiona jest także ciekawa analiza

jakościowa, w której dla zbioru ILSVRC12 jest przeanalizowane, jak wygląda sąsiedztwo kopii danego obrazka.



Powyższa ilustracja, zaczerpnięta z pracy doktorskiej, pokazuje na czerwono kopie punktów ze zbioru danych *Words Associations*.

Zanurzenia sąsiedztw w zastosowaniu do ewolucyjnego wyboru cech

W rozdziale 4 przedstawione są wyniki z prac

- Adrian Łańcucki, Indrajit Saha, Piotr Lipinski, *A new evolutionary gene selection technique*.
- Adrian Łańcucki, Indrajit Saha, Shib Sankar Bhowmick, Ujjwal Maulik, Piotr Lipiński, *A new evolutionary microRNA marker selection using next-generation sequencing data.*

zaprezentowanych na konferencji Congress on Evolutionary Computation odpowiednio w roku 2015 i 2016.

W obydwu pracach tematem są medyczne zbiory danych w postaci wektorowej charakteryzujące się bardzo małą liczbą przykładów (od 42 do 2280) i relatywnie dużą liczbą cech (od 199 do 2202). Celem prac było skonstruowanie klasyfikatorów dla powyższych zbiorów danych.

Przy konstrukcji klasyfikatorów zanurzenia sąsiedztwa są użyte jako fragment szerszego algorytmu optymalizacyjnego, w którym zewnętrzna optymalizacje jest dokonywana przy pomocy algorytmu ewolucyjnego, natomiast wewnętrzna optymalizacja przy pomocy metody SVM. Zanurzenia sąsiedztwa są użyte jako wstępne przekształcenie, w najprostszym wariancie polegające na włożeniu oryginalnych wektorów w zbiorze danych w przestrzeń jednowymiarową. Bez wstępnej redukcji wymiaru nie było realistyczne zastosowanie metod ewolucyjnych.

W pracy zebrane są eksperymenty wykonane przy użyciu innych algorytmów klasyfikacyjnych, w szczególności w pierwszej z powyższych prac, zaprezentowanej w rozdziale 4.2, w tabeli 4.3 można znaleźć zestawienie z klasyfikatorem SVM. We wszystkich badanych przypadkach połączenie t-SNE z SVM i algorytmem ewolucyjnym daje lepsze efekty niż uprzednio stosowane metody. W pracy *A new evolutionary microRNA marker selection using next-generation sequencing data* dodatkowo są ze sobą porównane dwa różne algorytmy ewolucyjne. Rezultaty dla obydwu algorytmów ewolucyjnych są zbliżone z przewagą dla wersji bazującej na Covariance Matrix Adaptation Evolution Strategy (CMA-ES).

Implementacja

Do pracy *Multipoint Neighbor Embedding* jest dostępna implementacja

https://github.com/alancucki/multipoint_tsne

o łącznej objętości 3782 linii.

Ocena pracy i konkluzje

Niniejsza praca ukazuje biegłość autora na wielu polach:

- podany jest nowy algorytm budujący wielopunktowe włożenia sąsiedztw,
- przeprowadzone są obszerne i dobrze skonstruowane eksperymenty,
- w rozdział 2-4, a także w dodatku A, w którym zestawione są funkcje kosztu dla t-SNE oraz dla włożeń word2vec i GloVe, autor wykazuje się zrozumieniem wyrafinowanego matematycznego aparatu i zdolnością samodzielnego poruszania się, jeśli chodzi o pojęcia współczesnej statystyki.

Prace pokazują szerokie zastosowania różnych wariantów metody t-SNE, przy czym na wybranych zbiorach danych medycznych osiągane są bardzo dobre wyniki eksperymentalne. Rozprawa jest starannie skonstruowana. W rozdziałach 2 i 3 podane jest wiele trafionych ilustracji włożeń, w rozdziale 4 dobrze dobrane są ukazane tam dane tabelaryczne. Dwie prace wchodzące w skład rozprawy były przedstawione na konferencji Congress on Evolutionary Computation, która jest jedną z głównych konferencji z zakresu algorytmów ewolucyjnych.

Stwierdzam, że praca pana Adriana Łańcuckiego spełnia ustawowe i zwyczajowe wymagania stawiane rozprawom doktorskim i wnoszę o dopuszczenie doktoranta do dalszych etapów przewodu doktorskiego.

Henryk Michalewski
Warszawa, 4 marca 2019