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

Poznań, 23.10.2018

Recenzja rozprawy doktorskiej mgr inż. Adriana Łańcuckiego pt. "Neighbor Embedding in Feature Selection and Multipoint Extensions"

1 Tematyka rozprawy

Tematem rozprawy mgr inż. Adriana Łańcuckiego są algorytmy redukcji wymiarowości, zwłaszcza te pracujące w trybie uczenia nienadzorowanym. Celem redukcji wymiarowości jest 'kompresja' (potencjalnie wysokowymiarowej) przestrzeni oryginalnych danych do przestrzeni o niższej liczbie wymiarów (często bardzo niskiej, np. 2 lub 3, zwłaszcza gdy celem jest wizualizacja danych, co często ma miejsce), przy jednocześnie możliwie wiernym zachowaniu struktury analizowanych danych (tj. wzajemnych relacji pomiędzy poszczególnymi obserwacjami, zwłaszcza w sensie relacji sąsiedztwa). Wobec lawinowo obecnie rosnących wolumenów danych, czyli tzw. powodzi danych (ang. *data deluge*), która przejawia się zarówno w rosnącej liczbie obserwacji jak i zwiększającej się liczbie zbieranych atrybutów/zmiennych (czyli właśnie wymiarów), możliwość wiernego i efektywnego odzwierciedlenia danych w przystępnej liczbie wymiarów ma istotne znaczenie praktyczne. W związku z tym tematykę pracy mgra Łańcuckiego uważam za aktualną, i zdecydowanie umiejscowioną w dyscyplinie informatyka, a dokładniej w nośnym obecnie nurcie (inteligentnej) analizy danych, odkrywania wiedzy i uczenia maszynowego.

2 Ocena treści rozprawy i wkładu oryginalnego

Rozprawa składa się z 5 rozdziałów, dwóch załączników, bibliografii, ma w sumie 90 stron, i jest bogato ilustrowana rysunkami i tabelami.

Mgr Łańcucki skupił się w rozprawie na wybranej i często obecnie stosowanej klasie algorytmów redukcji wymiarowości przez zanurzenie (ang. *embedding*), a dokładniej sto-

chastyczne zanurzenie sąsiedztw (ang. *stochastic neighborhood embedding*, SNE). W metodach tych odwzorowanie z oryginalnej do docelowej przestrzeni konstruuje się poprzez próbę reprodukcji w docelowej przestrzeni lokalnych rozkładów prawdopodobieństwa charakteryzujących sąsiedztwa pojedynczych obserwacji. Autor wskazuje iż, podobnie jak w innych technikach redukcji wymiarowości, zachowanie relacji pomiędzy obserwacjami jest trudne lub nawet niemożliwe (w ogólności d-wymiarowej przestrzeni da się umieścić jedynie $d + 1$ równoodległych od siebie parami obserwacji, a zatem redukcja d czyni to niemożliwym).

Rozdział 2 przedstawia przegląd literatury zanurzeń sąsiedztw. Przegląd jest szeroki, ale jednocześnie kompetentny, z klarownym podkreśleniem różnic pomiędzy poszczególnymi podejściami, co zdradza dobrą znajomość tematyki. Poziom zagłębienia się w temat jest odpowiedni, tj. Autor z jednej strony unika nadmiaru szczegółów, a z drugiej powierzchowności, oddając w tekście i formalizmach matematyczno-algorytmicznych istotę poszczególnych metod. Omawiane są zarówno metody o rodowodzie czysto statystycznym czy algebraicznym, jak i te oparte na sztucznych sieciach neuronowych. Interesujące jest także poświęcenie osobnej sekcji na wskazanie równoważności niektórych par metod (sekcja 2.2.6). Kwestionowałbym jedynie ideę równoległego prowadzenia dyskursu dla metod nadzorowanych (*feature extraction/construction/engineering*) oraz nienadzorowanych (*dimensionality reduction*), bo sformułowania tych zadań i stawiane w nich cele są jednak diametralnie różne.

Głównym oryginalnym przyczynkiem Autora, zaprezentowanym w rozdziale 3 rozprawy, jest duplikowanie obserwacji w docelowej przestrzeni, co pozwala uniknąć (lub osłabić znaczenie) problemów napotykanych przy próbie zachowania 'topologicznych' właściwości badanego zbioru danych. W tym celu zaproponował on metodę *multipoint t-SNE*, którą zaprezentował w rozdziale 3 rozprawy. Metoda ta iteracyjnie konstruuje zanurzenie, startując z zanurzenia losowego, w którym to procesie duplikuje wybrane obserwacje i przemieszcza powstałe w ten sposób duplikaty względem siebie według heurystyki bazującej na metaforze sił (sprężyn) oddziałujących na obserwacje, a pochodzących z obserwacji sąsiednich. Zaproponowane podejście (a właściwie rozszerzenie pewnego wariantu podejścia SNE, tj. algorytmu t-SNE) jest interesujące, nietrywialne i wydaje się dobrze uzasadnione.

Rozdział 3 prezentuje też ewaluację eksperymentalną proponowanego algorytmu. Doktorant przeprowadził ją na stosunkowo wielu zróżnicowanych zbiorach danych (ILSVRC12, COIL-20, NIPS co-autorships, Word2vec) o zdywersyfikowanej charakterystykach i rodowodach (obrazy, grafy współautorstwa, relacje podobieństwa słów). Mimo pewnych niedoskonałości prezentacji wyników (zob. dalsza część recenzji), ogólna wymowa rezultatów przemawia za zaproponowaną przez autora metodą; w szczególności, multipoint t-SNE systematycznie osiąga wyższe wartości miary precyzji na niż zwykłe t-SNE oraz inne warianty metody (np. rys. 3.5).

Rozdział 4 prezentuje metody które wykorzystują podejście t-SNE do projektowania reprezentacji rozwiązań kandydackich w ewolucyjnej selekcji cech z pomocą algorytmów ewolucyjnych, dokładniej ewolucji różnicowej (ang. *differential evolution*, sekcja 4.2), dobrze znanej i sprawdzonej metody ewolucyjnej adaptacji macierzy kowariancji (ang.

covariance matrix adaptation evolutionary strategy, CMA-ES sekcja 4.3). Podejście to oceniam jako oryginalne i ciekawe: zanurzanie sąsiedztw jest tu wykorzystane do uporządkowania (przesortowania) cech (których porządek w oryginalnym zbiorze danych jest w ogólności arbitralny) w taki sposób aby następnie pracujący na tym zbiorze ewolucyjny algorytm selekcji cech mógł modyfikować rozwiązania (genotypy) tak aby przekładały się one na selekcje cech o podobnych charakterystykach. Mimo pewnych niedociągnięć prezentacji i uchybień metodologicznych (zob. uwagi szczegółowe poniżej), ogólna wymowa eksperymentów (których przedmiotem jest selekcja atrybutów w zbiorach danych pochodzących z mikromacierzy) jest przekonująca: proponowane metody (t-SNE-DE-SVM dla ewolucji różnicowej oraz tSCES jako wariant CMA-ES) systematycznie zajmują czołowe miejsca w rankingach metod na rozważanych benchmarkach (sześciu w pierwszym przypadku (Tabela 4.1) oraz ośmiu w drugim (Tabela 4.6)). Mgr Łańcucki zamknął ten rozdział propozycją uogólnienia metody tSCES na więcej niż jednowymiarowe osadzenia cech (sekcja 4.4), gdzie w szczególności zademonstrował przydatność wielopunktowego rozszerzenia multipoint t-SNE zaproponowanego wcześniej w rozdziale 3.

Lektura pracy skłoniła mnie do pewnych uwag polemicznych, z których większość dotyczy prezentacji treści rozprawy:

- Rozdział 1 dobrze zarysowuje obszar i kontekst badań, ale moim zdaniem nie definiuje hipotezy badawczej i celów wystarczająco precyzyjnie. Co prawda ostatni akapit na s. 1 zarysowuje cel badań, ale robi to bardzo ogólnikowo („*The goal of this thesis is to propose a method, that recognizes and duplicates points, in order to construct meaningful embeddings.*"), i cel ten nie jest już dalej rozwinięty/przybliżony w tym rozdziale.

- Autor słusznie kojarzy (np. na początku sekcji 2.3) metody redukcji wymiarowości głównie z uczeniem nienadzorowanym, a metody selekcji cech głównie z uczeniem nadzorowanym. Wydaje się jednak nie wyjaśniać dlaczego tak jest – mianowicie ponieważ metody selekcji cech zmuszone są niejako do oceny 'jakości' pojedynczych cech lub ich podzbiorów, a naturalną miarą tak rozumianej jakości jest zdolność predykcyjna. Z drugiej strony warto może było zaznaczyć że metodologicznie rzecz biorąc selekcja cech jest szczególnym przypadkiem redukcji wymiarowości, w którym $X' \subset X$.

- Metoda Autora bazuje na metodzie t-SNE, która działa w sposób nienadzorowany; dlaczego zatem w Rys. 3.2 w sekcji 3.1.3 opisującej podejście Autora (multipoint t-SNE) pojawia się wzmianka o przykładach pozytywnych i negatywnych? W powiązaniu z tym, na s. 37 autor interpretuje wybór obserwacji do replikacji w kategoriach 'sił' pochodzących od przykładów pozytywnych i negatywnych, kojarząc te pierwsze z p_{ij} a drugie z q_{ij}. Niemniej p_{ij} to prawdopodobieństwa skojarzone z oryginalną przestrzenią, a q_{ij} z docelową (tak wielkości te zostały zdefiniowane w sekcji 2.2.3). Czyżby miało to oznaczać że autor doszukuje się 'sił przyciągających' w oryginalnej przestrzeni, a odpychających jedynie w tej docelowej?

- Prezentacja wyników w sekcjach eksperymentalnych, zwłaszcza w sekcji 3.3, jest

moim zdaniem niedopracowana: choć zwarta, jest też miejscami pobieżna i nieco chaotyczna: tekst nagle przełącza się z jednego zbioru danych na inny, zawiera niedociągnięcia i braki. Na przykład ostatnie zdanie na dole s. 42 zapowiada prezentację wyników w kategoriach miary NPR (*Neighborhood Preservation Ratio*, wprowadzonej wcześniej w pracy), a jedyne prezentowane wyniki ilościowe (w postaci wykresów) to precyzja (*precision*). Opisy zbiorów danych są raczej pobieżne, np. opisując zbiór NIPS autor pisze „*describes words similarities judged by human volunteers*", ale nie jest jasne jak owe 'similarities' zostały przełożone na zawartość macierzy relacji („square matrices of pairwise relations"). Miejscami opis parametryzacji metod i sposobu przeprowadzenia eksperymentu następuje po omówieniu wyników (np. trzy ostatnie akapity sekcji 3.3.2).
Niedopracowanie prezentacji wyników empirycznych uważam za niefortunne, bo nie odzwierciedla ona przez to w pełni wkładu koncepcyjnego autora (metoda multipoint t-SNE i jej skuteczność) ani znacznego (jak się spodziewam, sądząc po liczbie zaangażowanych zbiorów danych) nakładu pracy na przeprowadzenie eksperymentów.

- Autor nadinterpretuje wynik testu Friedmana (s. 62 i Tabela 4.1). Wynik ten istotnie pozwala odrzucić hipotezę zerową, ale hipotezą alternatywną w tym teście jest „przynajmniej jedna z metod jest statystycznie istotnie różna od innej". Jednak aby przekonać się o tym która z metod ma tę właściwość, należy przeprowadzić analizę *post-hoc*.

- Zestawienie Stochastic Neighborhood Embedding ze współczesnymi zanurzeniami słów (Word2Vec oraz GloVe) jest pouczające, i jako takie zasługiwałoby na włączenie w główny nurt pracy, a nie umieszczanie ich w załączniku (Appendix A). Jest to tym bardziej dziwne że autor sam deklaruje że wyniki te nie były wcześniej publikowane (co oczywiście z drugiej strony nie implikuje koniecznie ich oryginalności).

Z drobniejszych uwag polemicznych, prezentowanie metody MDS w sekcji 2.2.1 *Linear Dimensionality Reduction Methods* jest dyskusyjne, jako że MDS jest w ogólności metodą nieliniową, co klarownie wynika też z rysunku 2.1. Metody forward selection (dół strony 28) opisywane są tak jakby ich stosowanie ograniczone było do klasyfikatorów/metod liniowych, podczas gdy są one jednak ogólniejsze (co jednak autor wydaje się wiedzieć, wnosząc z wcześniejszych wzmianek w sekcji 2.3.2 Wrapper Feature Selection (s. 27)). Uwaga związana z przypisem dolnym na s. 6 jest interesująca („*this thesis adopts a less common notation for cross entropy Hp(q), what emphasizes this fact and cleanly distinguishes it from notation conflicts with joint entropy*"), choć zauważmy że generalnie w notacjach formalnych nie zakładamy że zapis $f(x, y)$ implikuje że f jest funkcją symetryczną. Skrót NPR (ang. *Neighborhood Preservation Ratio*) wprowadzony na s. 34 został rozwinięty dopiero na s. 42. Z kolei wsteczny odnośnik na str. 42 wskazuje s. 34 jako miejsce zdefiniowania NPR, i dopiero to pozwoliło mi domyśleć się że poprzedzający akapit na s. 34 to właśnie definicja NPR (choć nadal definicja ta ma charakter tekstowy,

nieformalny – niemniej jest zrozumiała). Tabele i rysunki są często umieszczone w niezbyt logicznych miejscach: np. Tabela 4.1 prezentuje zupełnie finalne wyniki (rangi i test Friedmana) jeszcze przed prezentacją ustawień parametrów, i na cztery strony przed pierwszym odwołaniem do tej tabeli. Symbol lr w Algorytmie 3.1 zamienia się na lr_i w opisie (choć z kontekstu da się to zrozumieć). Wydaje się że w definicjach pochodnej z funkcji min (3.4) można by z zachowaniem semantyki włączyć przypadek 3 do 1 w lewym równaniu (definicji), a przypadek 3 do 2 w prawym równaniu. W poprzedzającym opisie warto by przytoczyć ponownie definicję funkcji kosztu C, bo wymaga to obecnie powrotu do sekcji 2.2.3. Autorowi nie udało się też uniknąć pewnych powtórzeń; np. dolny akapit na s. 55 powtarza wiele elementów za poprzednim. Choć trudno oczekiwać aby praca była w pełni zamknięta/samowystarczalna (ang. self-contained), to pewne terminy można było trochę precyzyjniej przybliżyć (np. przestrzeń Hausdorffa na s. 8).

Zaznaczmy jednak że zdecydowana większość część argumentacji i wywodów w rozprawie prowadzona jest w przekonujący sposób. Autorowi udaje się też w większości przypadków unikać zbędnych powtórzeń, dzięki czemu praca jest zwarta a jednocześnie treściwa. Autor konsekwentnie używa ustalonej wcześniej przez siebie terminologii. Redakcja językowa i stylistyczna pracy jest staranna, z dobrym poziomem języka angielskiego; natrafiłem na zaledwie kilka błędów językowych, gramatycznych i typograficznych (np. zdanie „In comparison to ..." na s. 16), brak kropki na końcu drugiego akapitu na s. 8, brak „with" w „often measured ... the Vapnik-Chervonenkis dimension" na s. 53.

W ramach komentarzy/sugestii, w obu podejściach ewolucyjnych opisywanych w Rozdziale 4 autor realizuje de facto podejście dwukryterialne (drugi paragraf na s. 56 oraz formuła 4.15). Ciekawe byłoby wykorzystanie w miejsce tych zabiegów metod wielokryterialnej selekcji, np. NSGA-II lub NSGA-III [Deb et al.].

3 Konkluzja końcowa

Przedstawiona do oceny rozprawa doktorska mgr inż. Adriana Łańcuckiego zawiera oryginalne i wartościowe osiągnięcia, mocno podparte wynikami empirycznymi uzyskanymi na wymagających danych rzeczywistych, które stanowią znaczący przyczynek do zanurzeń sąsiedztw, a w szerszym kontekście do metod redukcji wymiarowości i selekcji cech, w tym tych najbardziej obecnie popularnych i studiowanych. Wymienione powyżej uwagi polemiczne odnośnie treści i prezentacji pracy nie podważają głównych konkluzji rozprawy i mojej pozytywnej jej oceny, zwłaszcza od strony przyczynków koncepcyjnych. Szczególnie przekonują mnie: (i) dogłębna znajomość tematyki, (ii) dobrze umotywowany i oryginalny charakter rozwinięcia *multipoint* przez 'klonowanie' obserwacji, (iii) oryginalny pomysł przesortowywania cech przy użyciu metod t-SNE na potrzeby ewolucyjnych algorytmów ich selekcji, oraz (iv) uniwersalność zaproponowanego podejścia, którą Autor zademonstrował na danych o bardzo zróżnicowanej charakterystyce i pochodzeniu: obrazach, grafach współautorstwa, grafach podobieństwa znaczenia słów, oraz kilku zbiorach danych mikromacierzowych. Uważam że cele postawione przez Autora pracy zostały osiągnięte.

Wobec powyższego stwierdzam, że **rozprawa doktorska mgr inż. Adriana Łań-**

cuckiego spełnia z nawiązką warunki stawiane przez ustawę o tytule naukowym i stopniach naukowych w odniesieniu do rozpraw doktorskich, a zatem powinna być dopuszczona do publicznej obrony, o co wnoszę do Rady Wydziału Matematyki i Informatyki Uniwersytetu Wrocławskiego.