prof. dr hab. Jan Mielniczuk
Instytut Podstaw Informatyki PAN
i Wydział MiNI PW

25.IV.2019

Recenzja rozprawy doktorskiej mgr Krzysztofa Gogolewskiego

Matrix methods in transcriptomic and metabolomic data analysis

Omówienie rozprawy

Rozprawa składa się z trzech głównych części dotyczących różnych problemów analizy danych transkryptomicznych i metabolomicznych.

W częsci pierwszej rozpatruje się problem detekcji heterogeniczności populacji komórek na podstawie profili transkryptomicznych. W tym celu dokonuje się przybliżonej dekompozycji sygnału transkryptomicznego opisanego macierzą $A = (A_{ij})$, gdzie A_{ij} oznacza aktywność genu i w próbce j na wkłady z K procesów, przy czym aktywność genu i związana z procesem l wynosi W_{il}, a l-ty proces jest reprezentowany w próbce j w proporcji $H_{l,j}$. Wykorzystuje się w tym celu metodę Approx NMF (Approximative Nonnegative Matrix Factorisation) dekompozycji macierzy A na sumę iloczynu dwóch nieujemnych macierzy W i H i macierz szumu (Brunet et al, 2004). Po tym kroku dokonuje się wyboru biomarkerów najbardziej różnicujących procesy, które następnie wykorzystuje się do końcowej dekompozycji sygnału transkryptomicznego (metoda nazwana Molecular Process Heterogenieity, MPH). Algorytm stosuje się do linii komórkowej neuroblastomy w trzech warunkach eksperymentalnych: kontrolnych, w eksperymencie oddziaływania na komórki C2-ceramidem i C2-ceramidem łacznie z inhibitorem PARP.

Druga część rozprawy dotyczy estymacji niskorzędowego sygnału transkryptomicznego. Rozpatruje się macierz $A = (a_{ij})$, której element a_{ij} jest teraz liczbą genów i w komórce j i przyjmuje się, że macierz A dekomponuje się na sumę $L + S$ lub na sumę $L + S + E$, gdzie macierz L jest macierzą niskiego rzędu sygnału transkryptomicznego, S jest macierzą rzadką, a E macierzą szumu. Dekompozycji poszukuje się szukając minimum funkcji kosztu

$$||L||_* + \lambda_1 ||S||_1, \text{ gdy } A = L + S \tag{1}$$

w pierwszym przypadku, i funkcji kosztu

$$||L||_* + \lambda_1||S||_1 + \lambda_2||E||_F, \text{ gdy } A = L + S + E \tag{2}$$

w przypadku drugim, gdzie $||\cdot||_*, ||\cdot||_1, ||\cdot||_F$ są odpowiednio normą nuklearną, normą L^1 i norma L^2 (Frobeniusa) macierzy.

W pierwszym problemie dokonuje się modyfikacji adaptowanego do tego przypadku algorytmu rozszerzonego Lagrangianu AGL zaproponowanego przez Yuana i Yanga (2009), polegającej na wprowadzeniu zmiennego progu miękkiego obcinania oraz parametru k_0, będącego ograniczeniem rzędu macierzy L. Metoda ta została nazwana w pracy metodą obciętego odpornego PCA: trRPCA. W drugim przypadku postępowanie jest podobne i jego celem jest znalezienie minimum ze względu na L, S i E operatora AGL

$$l(L, S, E, Y) = ||L||_* + \lambda_1||S||_1 + \lambda_2||E||_F + < Y, A - L - S - E > + \mu||A - L - S - E||_F^2,$$

gdzie Y jest macierzą Lagrangianu, a $< A, B >= tr(AB^T)$. Modyfikacja metody AGL w tym przypadku prowadzi do algorytmu trPCAL2, przy czym w odróżnieniu od macierzy S i L, których kolejne aproksymacje dostaje się przez miękkie obcinanie (elementów macierzy S lub dekompozycji SVD macierzy L odpowiednio), kolejne aproksymacje macierzy szumu E dostaje się przez zastosowanie przekształcenia $\mathcal{E}$, którego postać wyprowadza się w pracy. Użyteczność metody trPCAL2 sprawdza się na komórkach PBMC (Peripherial Blood Mononuclear Cells): na otrzymanej macierzy L wykonuje się hierarchiczną analizę skupień i porównując jej zgodność z podziałem otrzymanym metodą korelacyjną.

Trzeci rozpatrywany problem (rozdział czwarty rozprawy) dotyczy integracji danych transkryptomicznych z modelem metabolizmu ludzkiego RECON 2.2. Pokazano, że biomarkery metabolomiczne, które wykrywa się w takim procesie stosując metodę stosowaną w literaturze (Shlomi et al, 2008) mogą być fałszywymi sygnałami, gdyż pozornie istotne sygnały można uzyskać stosując te metody na losowym zbiorze danych. Z tego względu w pracy proponuje się metodę likwidacji obciążenia, która pozwala uniknąć problemu przeuczenia. W rozdziale 5 rozprawy przedstawiono badania dotyczące roli liczby kopii genu FOXF1 na regulację jego aktywności.

Ocena rozprawy

Metoda NMF była wykorzystywana do analizy sygnału transkryptomicznego dla danych pochodzacych z różnych źródeł, o czym pisze Autor na str. 28 rozprawy. Nowa propozycja

omówiona w rozdziale 2 i w współautorskiej pracy Gogolewski et al., 2017 polega na dwustopniowości procedury MPH: w pierwszym etapie po wstępnej filtracji genów wykorzystując procedurę `DSection` dokonuje się wstępnej dekompozycji A i na jej podstawie wyboru istotnych biomarkerów, w drugim, przy użyciu procedury `ssKL` mającej podobny cel otrzymuje się ostateczną dekompozycję i końcowy wybór różnicujących genów. Wykorzystanie tej metody w analizie linii komórkowej neuroblastomy jest przekonujące i pokazuje, że na podstawie sygnału transkryptomicznego wybranych tą metodą genów można potwierdzić współwystępowanie tu dwóch procesów; w wyraźny sposób wskazuje na to macierz aktywności genów na rys. 2.2 w rozprawie. Pokazano również, że istotny wpływ na przebieg procesu ma wykorzystanie obok C2-ceramidu dodatkowo inhibitora.

Problem analizy sygnału transkryptomicznego rozpatrywany w rozdziale trzecim rozprawy, w szczególności przez zastosowanie metod nienadzorowanych, należy do ważnych problemów bioniformatyki (Lowe at al, 2017). Podstawowe osiągnięcia w tej dziedzinie są kompetentnie opisane przez mgr Gogolewskiego w podrozdziale 3.2. Metody wykorzystujące analizę składowych głównych (PCA) mają tutaj podstawowe znaczenie i naturalnym nowym pomysłem była próba wykorzystania odpornej metody PCA, RPCA (Candès et al, 2011), stosowanej dotychczas przede wszystkim do analizy sekwencji video, w kontekście nauk biologicznych. Istotnym osiągnięciem pracy jest zaproponowanie rozszerzonego modelu dekompozycji niskorangowej, uwzględniającego szum i algorytmu przybliżającego składniki tej dekompozycji. To podejście ma ważne implikacje metodologiczne: zamiast usiłować wykluczyć metodami preprocessingu dużą zmienność danych (do pewnego stopnia nieuniknioną) związaną w szczególności z techniką ich pozyskania stara się tę zmienność uwzględnić i oszacować w modelu.

Zastosowanie tej metody do sekwencji RNA dla pojedyńczych komórek (scRNA-seq jednojądrzastych komórek krwi obwodowej) dało rozbicie danych na 4 skupienia (i jednego z otrzymanych skupień na dodatkowe dwa skupienia) dobrze korespondujące z podziałem otrzymanym metoda korelacyjną. Rownież uzyskany podział na skupienia pokazuje jasno zmianę ekspresji genów CD14 i FCGR3A w skupieniach monocytów oraz koekspresję dwóch innych często badanych genów.

Wyniki te, obok propozycji metody MPH, uznaję za podstawowe osiągnięcie rozprawy doktorskiej: Doktorant przedstawił użyteczne rozszerzenie istniejącego modelu dekompozycji niskorangowej, algorytm przybliżonego rozwiązania oraz zastosował go z sukcesem do trud-

nych danych, radząc sobie w szczególności z problemem doboru parametrów regularyzacji λ_1 *i* λ_2.

W rozdziale 4 pokazano, że metoda tworzenia 'krajobrazów' metabolicznych oparta na optymalizacji parametrów modelu RECON 2.2 (Shlomi et al, 2008) może prowadzić do pozornych skupień (otrzymanych na podstawie kilku pierwszych składowych głównych rozwiązania). Efekt ten jest związany z nadreprezentacją pewnych specyficznych reguł genetycznych koniecznych dla zajścia odpowiedniej reakcji. Zaproponowano usunięcie tej wady poprzez agregację reakacji odpowiadających tej samej regule. Pokazano, że metoda jest skuteczna w przypadku zbioru danych dotyczących raka nerki. Zwrócenie uwagi na ten problem dowodzi dobrego zrozumienia przez mgr Gogolewskiego problemu przeuczenia, który często wystepuje w uczeniu maszynowym (por. Ioannidis, *Why most of the published research findings are false*, PLOS Medicine, 2005).

Słabszym elementem pracy jest jej nadmierna lapidarność, która nie pozwala do końca prześledzić motywacji piszącego oraz szczegółów implementacyjnych algorytmów oraz wyników analizy (kilka braków tego typu podaję w uwagach szczegółowych). W przypadku metody MPH nie znalazłem wyjasnienia, co wnosi drugi etap tej metody wykorzystujący `ssKL` i dlaczego tak skonstruowana procedura jest bardziej efektywna od dwukrotnie zastosowanej procedury `DSection`.

Dla odpornych metod składowych głównych nie jest jasne, co skłoniło Autora do modyfikacji algorytmu RPCA (rozumiem, że dla rozpatrywanych danych transkryptomicznych aproksymacja L nie była niskiego rzędu) i zastąpienia go algorytmem trPCA, podobna uwaga tyczy się przejścia od trPCA do trPCAL2. Własności obu modyfikacji, w szczególności ich zbieżność oraz zbieżność do minimów funkcji kryterialnych (??) i (??) są w doktoracie pominięte. *W przypadku rozprawy doktorskiej z informatyki w dziedzinie nauk matematycznych (bo do takiej dziedziny przypisana jest rozprawa) jest to istotny brak.* Analiza formalna tego problemu może być złożona, ale możliwa była analiza zbieżności w sztucznie skonstruowanych, ale reprezentatywnych modelach symulacyjnych.

Również podstawową kwestią, zupełnie w rozprawie pominiętą jest kwestia identyfikowalności dekompozycji $A = L + S + E$, której poszukujemy. Nie ułatwia sytuacji fakt, że Doktorant nie precyzuje, co rozumie pod pojęciem 'gęstego szumu', ograniczając się tylko do intuicji.

Choć obie propozycje modyfikacji RPCA, jak pisałem wyżej, są interesujące i rozwijają

nowe propozycje analizy danych wysokowymiarowych, uzasadnione jest również pytanie, jak mają sie one do do propozycji znacznie mniej złożonych, ale narzucających się w tym kontekście: przykładowo, zamiast poszukiwania dekompozycji $A = L + S + E$ można było zastosować jedną z metod odpornej anlizy skupień (np metoda obcinanych k-średnich lub metoda k-średnich wykorzystująca normę L^1, por. Gordon, *Robust cluster analysis and variable selection*) bezpośrednio na macierzy A.

Na koniec uwaga ogólna: głównym powodem nadmiernej lapidarności i braku istotnych szczegółów w rozprawie jest w moim przekonaniu zbyt duża liczba poruszanych w niej zagadnień (i nie zawsze związanych z jej tytułem, jak rozdział 5). W efekcie rozprawa jest niespójna. Szersze przedstawienie zagadnień rozdziału 2 i 3 z bardziej szczegółowym omówieniem zalet przedstawionych metod, przyniosłoby znacznie bardziej dojrzałą rozprawę doktorską.

Kilka uwag szczegółowych:

1. W przypadku metody MPH forma kosztu funkcji $\mathcal{F}(A, W, H)$ na str. 31 rozprawy pozostawiona jest bez żadnego komentarza. Nie dowiedziałem się tez nic na temat jakości dopasowania $\hat{W}\hat{H}$ do A.

2. Na str. 36 przydałby się komentarz, że wstępna filtracja ogranicza zbiór rozpatrywanych dalej genów do mających silne efekty główne, a pomija geny mające jedynie interakcyjny wpływ na ekspresję;

3. str. 58: nie wiadomo, przy użyciu jakiej metody stwierdzono zwiększoną aktywność genów CD8A i CD8B w jednym ze skupień i czy zastosowano ją indywidualnie czy łacznie do tych genów;

4. str. 54_3: nie wiem, co tutaj oznacza 'najbardziej stosowna liczba skupień';

5. str. 58: nie jest jasne, którą z metod hierarchicznej analizy skupień różniących sie miarami odmienności skupień zastosowano;

6. str. 51: rola parametru c w algorytmie trPCA nie jest dyskutowana, nie jest podana też jego wartość używana w eksperymentach;

Istotne literówki: k zamiast k_0 (50_1); μ_i zamiast μ_{i+1} (punkt 8 algorytmu 2 str. 51); λ_2 zamiast μ (53_5), k zamiast j w $L_k^{(i)}$ (72_5).

Sama rozprawa, jak i dorobek publikacyjny mgr K. Gogolewskiego (3 współautorskie prace dotyczace tematyki rozprawy, 4 współautorskie prace dotyczace innych badań, w tym jedna w *BMC Medical Genetics*) świadczą o jego intensywnym udziale w pracach zespołu

badawczego analizy danych trnskryptomicznych i metabolomicznych.
Mimo zastrzeżeń przedstawionych powyżej moja końcowa opinia jest zdecydowanie pozytywna.

Konkuzja

W mojej opinii rozprawa doktorska przedstawiona przez mgr Krzysztofa Gogolewskiego spełnia ustawowe wymogi stawiane rozprawom doktorskim w dziedzinie nauk matematycznych, dyscyplina informatyka i wnoszę o dopuszczenie go do dalszych etapów przewodu doktorskiego.

Jan Mielniczuk