Recenzja pracy doktorskiej mgr inż. Antoniny Krajewskiej

Efficient matrix completion for data recovery in data-driven IT applications

(02.11.2023)

1 Wstęp

Rozprawa doktorska mgr. inż. Antoniny Krajewskiej pt. „*Efficient matrix completion for data recovery in data-driven IT applications*" dotyczy zagadnienia uzupełniania macierzy, której elementy znamy jedynie częściowo. Jest to ważne i aktualne zagadnienie, a praca wymienia wiele jego przykładowych zastosowań.

Po omówieniu i analizie istniejących podejść, Doktorantka proponuje własne rozwiązanie: CSMC (*Column Selected Matrix Completion*). Dekomponuje ono zadanie na dwa etapy: wybór podzbioru kolumn (*Column Subset Selection*) oraz uzupełnienie powstałej w pierwszej fazie macierzy małego rzędu (*Low-rank Matrix Completion*). Dekompozycja taka pozwala na zwiększenie efektywności algorytmu, dzięki operowaniu na mniejszej macierzy.

Opracowana metoda zostaje przez Doktorantkę przeanalizowana zarówno pod względem własności teoretycznych, jak również pod względem praktycznym, poprzez przeprowadzenie szeregu eksperymentów numerycznych dla trzech zadań z realistycznymi danymi.

2 Omówienie dorobku Doktorantki

Dorobek Doktorantki obejmuje, poza recenzowaną rozprawą, kilka artykułów, opublikowanych w materiałach konferencji naukowych (HPCS) oraz czasopismach (Control and Cybernetics, Journal of Telecommunications and Information Technology). Są wśród nich artykuły napisane samodzielnie.

Dorobek jest, w mojej ocenie, ciekawy i wartościowy, a obejmuje także prace dość luźno związane z tematyką zaprezentowanej rozprawy doktorskiej.

Jest to dorobek odpowiedni do aplikowania o tytuł doktora – i to z wyraźnym nadmiarem.

3 Treść rozprawy doktorskiej

Praca składa się z dziesięciu rozdziałów.

Pierwszy z nich zwięźle przedstawia rozpatrywane zagadnienie oraz prezentuje dwie hipotezy badawcze, które – w pewnym uproszczeniu – można sformułować następująco:

- Jakość zaproponowanego algorytmu uzupełniania macierzy nie odbiega od jakości dotychczasowych algorytmów, zaproponowanych w literaturze.
- Wydajność oraz skalowalność zaproponowanego algorytmu jest lepsza (dosłownie: „konkurencyjna") względem owych starszych podejść.

Rozdział drugi również jest krótki, a przedstawia podstawowe pojęcia oraz używane oznaczenia.

W rozdziale trzecim precyzyjnie sformułowane zostaje zadanie uzupełnienia macierzy, zaprezentowana zostaje ogólna idea proponowanego algorytmu CSMC (podrozdział 3.4) oraz przedstawione zostaje uzasadnienie, dlaczego jest to ważne zadanie, z omówieniem przykładowych zastosowań.

W rozdziale czwartym Doktorantka dokonuje przeglądu istniejących algorytmów uzupełniania macierzy – dokładnych i niedokładnych – porównując ich zalety i wady.

Rozdział piąty skupia się na dwu metodach uzupełniania macierzy, stosujących optymalizację wypukłą tzw. normy jądrowej. Można by w tym momencie stwierdzić, że jest cokolwiek nielogiczne, iż rozdział piąty nie jest podrozdziałem czwartego. Jest to jednak uzasadnione zarówno jego rozmiarem, jak i znaczeniem jego treści dla badań Doktorantki. Powinno to być tylko jakoś skomentowane, np. pod koniec rozdziału czwartego mogłaby być informacja, że i dlaczego pewne metody zostaną omówione w oddzielnym rozdziale.

Rozdział szósty omawia zagadnienie wyboru podmacierzy, składającej się z kolumn macierzy oryginalnej, aby uzyskać jej aproksymację o niższym rzędzie (CSS – *Column Subset Selection*). Omówione tu metody są bardzo istotne dla działania algorytmu, opracowanego przez Doktorantkę. Rozdział jest, w porównaniu z innymi, nieco chaotyczny: wspomina o kilku algorytmach, nie podając ich szczegółów, omawia także dekompozycję CUR macierzy; ma to oczywiście swoje uzasadnienie, ale może być mylące.

Rozdział siódmy można nazwać najważniejszym – prezentuje on bowiem zaproponowaną metodą CSMC, czyli uzupełniania macierzy, z wykorzystaniem wyboru kolumn. Ogólne omówienie tego algorytmu znalazło się już wcześniej w podrozdziale 3.4, co chyba nie było najlepszym pomysłem na strukturę pracy; jest to jednakże drobiazg. W rozdziale siódmym przedstawione zostają trzy warianty metody CSMC:

- CSNN (*Column Selected Nuclear Norm*),
- CSPGD (*Column Selected Proximal Gradient Descent*) i
- CSPGD-adam – wariant poprzedniego rozwiązania, stosujący optymalizację typu „Adam".

Różnią się ona użytym algorytmem optymalizacji. Są to, odpowiednio: programowanie półokreślone (*semidefinite programming*), metoda PGD (*Proximal Gradient Descent*) oraz – oczywiście – optymalizacja Adam, zastosowana do zadania najmniejszych kwadratów.

Następnie omówiony zostaje schemat metody oraz jej własności teoretyczne. Mamy tu twierdzenia – zarówno cytowane z prac innych autorów,

jak i udowodnione przez Doktorantkę. Wydaje się, iż powinny one być jaśniej rozdzielone, uwypuklając oryginalne twierdzenia Autorki i ich dowody.

Rozdział ósmy przedstawia eksperymenty numeryczne. Wszystkie trzy wersje metody okazały się przydatne, zależnie od wymiaru zadania oraz jego własności. Część eksperymentalna wydaje się dobrze opisana.

Rozdział dziewiąty omawia przykłady zastosowań praktycznych. Dotyczą one, m.in., rekomendacji filmów (zadanie związane z Netflix Prize) oraz naprawy uszkodzonych obrazów. Zwłaszcza to drugie zastosowanie jest bardzo ciekawe i ilustruje siłę opracowanego przez Doktorantkę podejścia, w sposób przemawiający do wyobraźni.

Rozdział dziesiąty zawiera podsumowanie i wnioski. Uważam je za poprawne. Jako swoje osiągnięcia Doktorantka wymienia:

- opracowanie dwuetapowego algorytmu CSMC, przeznaczonego do uzupełniania „prostokątnych macierzy, dla których jeden z wymiarów jest istotnie większy od drugiego" (s. 119),
- opracowanie trzech, wymienionych powyżej, implementacji tego algorytmu,
- formalną analizę algorytmu CSMC, a zwłaszcza błędu drugiej fazy,
- opracowanie otwartej biblioteki w języku Python 3, implementującej owe trzy algorytmy – w wersji zarówno dla CPU, jak i GPU,
- utworzenie otwartej biblioteki z benchmarkami,
- analiza numeryczna i porównanie wydajności opracowanych metod ze znanymi z literatury.

Zgadzam się z treścią tej listy.

Bibliografia, która następuje po rozdziale 10, jest bardzo obszerna. Zawiera 197 publikacji i uważam ją za kompletną.

Ogólnie, układ pracy uważam – przy uwzględnieniu powyższych uwag – za poprawny. Mógłby być jeszcze lepszy, po wprowadzeniu tych drobnych zmian, ale w obecnej postaci, jest, jak najbardziej, akceptowalny.

4 Strona edycyjna pracy

Praca jest napisana w języku angielskim. Od strony edycyjnej jest napisana starannie. Język rozprawy uważam za bardzo poprawny; nie zauważyłem nawet niezręczności stylistycznych, zazwyczaj spotykanych u tak młodych badaczy.

Warto też podkreślić, że w pracy znajdują się listy użytych oznaczeń oraz skrótów – nie są one, w tym przypadku bardzo rozbudowane, ale ich wprowadzenie uważam za bardzo dobry zwyczaj, usprawniający niekiedy znacznie zapoznawanie się z treścią.

Więcej zastrzeżeń można mieć do edycji bibliografii. Pojawia się tutaj, częste u osób korzystających ze środowiska LaTeXoraz BibTeX, pomijanie wielkich liter w tytułach artykułów. Z małej litery pisane są nie tylko np.

nazwy rozkładów macierzy (np. „svd" w pozycjach [11, 12], czy „cur" w [33, 34], ale także inne skróty, np. „mri" w pozycji [23], „lstm" w [63], a nawet nazwiska, jak np. „frank-wolfe" w pozycji [78]). Jest to, jak wspomniałem, często spotykane, ale bardzo niewłaściwe i absolutnie nieuzasadnione kaleczenie języka; warto zatem zwrócić na tę praktykę uwagę.

Błędów w bibliografii jest więcej. Niektóre pozycje nie mają podanych autorów ([168], [189]), co jest ewidentnym niedopatrzeniem edycyjnym, gdyż są to zwyczajne artykuły, o jasno i jednoznacznie określonych autorach. Z kolei pozycja [159] nie podaje źródła.

Złym zwyczajem wydaje mi się także – ale tu na pewno można by się spierać – umieszczenie pozycji bibliograficznych w kolejności występowania w tekście, a nie alfabetycznie. Zwłaszcza przy tak obszernej bibliografii może to bardzo utrudnić wyszukiwanie pozycji bibliograficznych; bywa jednak stosowane przez niektórych badaczy. Osobiście pozostaję zwolennikiem stosowania kolejności alfabetycznej.

5 Oprogramowanie

Integralną częścią osiągnięcia doktorskiego mgr. inż. Antoniny Krajewskiej jest opracowane oprogramowanie, realizujące zaproponowane w pracy algorytmy. Oprogramowanie to zostało publicznie udostępnione za pomocą serwisu GitHub, co jest bardzo dobrą praktyką i uważam to za wielką zaletę pracy.

Warto w tym miejscu podkreślić, że oprogramowanie napisane zostało, przynajmniej w mojej ocenie, profesjonalnie, z wykorzystaniem dobrych praktyk programowania w języku Python. Mniej profesjonalna jest, niestety, dostępna dokumentacja, zawierająca kilka nieaktualnych bądź mylnych informacji, które mogą wprowadzić w błąd – zwłaszcza niedoświadczonego użytkownika, pragnącego skorzystać lub przetestować napisane przez Doktorantkę skrypty. Przykładowo, choć oprogramowanie udostępnione zostało finalnie (jak już wspomniano) w serwisie GitHub, w dokumentacji mowa jest o innym serwisie – GitLab (który, skądinąd, również byłby bardzo odpowiednim miejscem na udostępnienie kodów Doktorantki), a także o innym jeszcze serwisie, będącym (o ile rozumiem) wewnętrznym serwisem firmy NASK. Nie jest to bez znaczenia, gdyż dokumentacja (plik `README.md`) omawia sposób tworzenia np. *merge requests* i innych czynności, które dla GitHuba wyglądałyby nieco inaczej.

Można także zauważyć pewną niestosowność komentarzy, wyświetlanych np. przez funkcję `sgrad_ls`.

6 Podsumowanie

Pracę doktorską mgr inż. Antoniny Krajewskiej oceniam wysoko. Praca zawiera ciekawe, innowacyjne osiągnięcia i jest – pomimo drobnych, opisanych powyżej, wad – bardzo dobrze napisana. Praca spełnia wymagania stawiane pracom doktorskim – w moim pojęciu z wyraźnym nadmiarem. Wnioskuję

zatem o przyjęcie rzeczonej pracy i dopuszczenie Doktorantki do dalszych etapów przewodu doktorskiego.

Nie uważam się za wystarczająco doświadczonego, a także kompetentnego – ani w dziedzinie rachunku macierzy, ani w wąsko pojętej dziedzinie uczenia maszynowego – aby samodzielnie wnioskować o wyróżnienie; nie mam jednak zastrzeżeń, gdyby podobny wniosek wysunęli inni recenzenci.