Prof. Dr hab. inż. Ewaryst Rafajłowicz

Członek korespondent PAN

Wydział Informatyki i Telekomunikacji

Politechnika Wrocławska

Recenzja rozprawy doktorskiej

mgr Antoniny Krajewskiej

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

Niniejsza recenzja została napisana na zlecenie Rady Naukowej Instytutu Badań Systemowych PAN, w związku z z toczącym się przewodem doktorskim Pani mgr Antoniny Krajewskiej.

Tematyka rozprawy

Recenzowana rozprawa dotyczy problematyki uzupełniania macierzy w oparciu o pewien procent jej elementów, które są dostępne oraz o informację dodatkową, dotyczącą całej klasy rozpatrywanych macierzy. Problemy tego rodzaju są w ostatnich latach intensywnie badane w wielu ośrodkach akademickich z użyciem różnych podejść. Zainteresowanie nimi wynika z faktu, że wraz z gromadzeniem coraz większej liczby danych, coraz częściej okazuje się iż są one nie w pełni kompletne, a przez to mniej użyteczne. Oprócz zastosowań badanych przez Doktorantkę, wskazać można cały obszar danych medycznych, gdzie niekompletne dane utrudniają stosowanie metod klasteryzacji, klasyfikacji i głębokiego uczenia. Ponadto, jak wskazuje mgr A. Krajewska, uzupełnianie macierzy o brakujące kolumny może mieć zdolności predykcji, które są dodatkowo interesujące, gdyż wiedza a priori wyrażana jest tutaj w terminach innych niż w podejściach probabilistycznych.

Doktorantka udokumentowała w swojej rozprawie fakt, że istniejące algorytmy uzupełniania macierzy charakteryzują się bardzo dużymi nakładami obliczeniowymi. Zatem, każde ich zmniejszenie, przy zachowaniu porównywalnej dokładności, jest istotne. Takie motywacje powodują, że ***tematykę rozprawy uważam za aktualną i ważną.***

We Wstępie mgr A. Krajewska sformułowała i omówiła dwie hipotezy badawcze, które można zinterpretować jako tezę rozprawy. W swobodnym przekładzie można sformułować ją następująco:

Jakość rozwiązań proponowanej metody (podejścia) jest porównywalna z metodami uzupełniania macierzy opisanymi w literaturze, a ważnym celem pracy jest pokazanie, że możliwe jest uzyskiwanie rozwiązań o teoretycznie gwarantowanej jakości. Implementacje tej metody są efektywne, skalowalne i konkurencyjne wobec innych metod, które bazują na minimalizacji normy nuklearnej macierzy.

Teza rozprawy sformułowana została w sposób przekonywujący i weryfikowalny.

Zawartość i kompozycja rozprawy

> Zanim przejdę do omówienia zawartości rozprawy, potrzebna jest pewna uwaga terminologiczna. Na określenie swego osiągnięcia Autorka używa terminu*: Column Selected Matrix Completion* wraz z określeniem metoda lub podejście (*approach*). W moim odczuciu opracowała Ona szersze podejście, które zbadała i skonkretyzowała w postaci trzech istotnie różnych algorytmów. Dlatego dalej będę używał terminu „podejście CSMC", korzystając z oryginalnego akronimu Autorki. Podobną konwencję zastosuję do wariantów podejścia CSMC, które są rozważane w rozprawie.

Rozprawa liczy 136 stron i składa się z 10 rozdziałów oraz obszernej, liczącej 197 pozycji, bibliografii. Zawiera także spisy: treści, oznaczeń, akronimów, tabel oraz rysunków.

Przedstawiając w skrócie zawartość poszczególnych rozdziałów będę jednocześnie zwracał uwagę na osiągnięcia mgr A. Krajewskiej. Już na tym etapie warto zaznaczyć, że rozdziały od 1 do 6 mają jednocześnie charakter bardzo kompetentnego przeglądu literatury z jednoczesnym wprowadzeniem Czytelnika w podstawowe fakty potrzebne do zrozumienia koncepcji proponowanego podejścia CSMC. Rozdziały te liczą łącznie 50 stron, natomiast rozdziały od 7 do 10, które zawierają zasadniczy wkład mgr A. Krajewskiej, mają objętość około 70 stron.

Oprócz wspomnianego już opisu hipotez badawczych, rozdział wstępny zawiera także objaśnienie struktury dysertacji. Wstępny charakter ma także Rozdział 2, w którym znajdujemy definicje i przyjętą notację. Ważny dla zrozumienia motywacji i zamysłów Autorki jest Rozdział 3, który przedstawia ścisłe sformułowanie problemów rozważanych w rozprawie na tle zagadnień uzupełniania macierzy niskiego rzędu, ich zastosowań oraz ich związków z próbkowaniem oszczędnym. (*compressed sensing*).

Wprowadzenie do metod uzupełniania macierzy niskiego rzędu zawiera Rozdział 4, w którym Autorka dużo uwagi poświęca metodom bazującym na programowaniu wypukłym, gdyż również w Jej podejściu algorytmy programowania wypukłego odgrywają istotną rolę. Przegląd zawiera także metody uzupełniania macierzy niskiego rzędu w oparciu o faktoryzację macierzy i krótkie podsumowanie innych podejść.

Rozdział 5 jest istotny, gdyż daje przegląd pojęć i rezultatów związanych z metodami uzupełniania macierzy w wyniku minimalizacji nuklearnej normy różnicy: macierzy częściowo znanej i macierzy przybliżającej. Jest to ważny element konstrukcji podejścia proponowanego przez Autorkę. Rozdział ten zawiera także przegląd algorytmów programowania półokreślonego i gradientowych, używanych w kompletowaniu macierzy. Najważniejsza część tego rozdziału, to omówienie podstawowych pojęć i rezultatów związanych ze współczynnikiem koherencji macierzy, które są dalej podstawą do ważnych rezultatów Autorki, dotyczących możliwości rekonstrukcji macierzy, przy założeniach o jej współczynniku koherencji i liczbie znanych kolumn.

W Rozdziale 6, mgr A. Krajewska dokonuje przeglądu metod selekcji podzbioru kolumn macierzy (CSS) oraz wyboru kolumn i wierszy macierzy (CUR). Również ten rozdział zawiera bardzo staranny przegląd literatury i jest istotny z punktu widzenia podejścia proponowanego przez Doktorantkę.

Szczegółowy opis i badania teoretyczne proponowanego, nowego podejścia CSMC zawarte są w Rozdziale 7. Podejście to składa się z dwóch kroków:

I. W pierwszym z nich, losowany jest podzbiór kolumn macierzy, której elementy mają zostać uzupełnione. Tak uzyskana macierz jest uzupełniana za pomocą wybranego algorytmu rozwiązującego to zadanie.

II. W drugim kroku, rozwiązywane jest zadanie minimalizacji sumy kwadratów odchyleń, które obliczane są w oparciu o znane elementy rekonstruowanej macierzy oraz o elementy uzupełnione w pierwszym etapie selekcji kolumn.

Powyższe podejście jest bardzo przemyślaną i trafną kombinacją znanych problemów, co Autorka szeroko dokumentuje w dalszej części rozprawy. Każdy z tych problemów można rozwiązywać za pomocą co najmniej kilku znanych lub nowych metod, co w istocie otwiera całkiem nowy nurt badawczy. W celu uzasadnienia podejścia CSMC, Doktorantka wybrała trzy konkretne metody, które zaimplementowała i poddała szerokim badaniom porównawczym. Są to:

1. Column Selected Nuclear Norm (CSNN) – w algorytmie tym minimalizowana jest norma nuklearna, a do możliwie dokładnej minimalizacji używa się programowania półokreślonego.

2. Column Selected Proximal Gradient Descent (CSPGD) – to algorytm przybliżonej

minimalizacji normy nuklearnej z zastosowaniem przybliżonego gradientu.

3. CSPGD-adam to wersja Algorytmu CSPGD, w której do minimalizacji sumy kwadratów odchyleń użyto znanej metody o akronimie ADAM.

Algorytmy te zostały dokładnie opisane w postaci pseudokodów, ze wskazaniem szczegółowych algorytmów używanych w krokach I i II i rozmiarów macierzy uzupełnianej, dla których są one dedykowane.

Z punktu widzenia wskazania warunków dostatecznych, przy których proponowane podejście zapewnia, z zadanym prawdopodobieństwem bliskim 1, dokładną rekonstrukcję macierzy, zawarte są w podrozdziale 7.2. Warunki te sformułowała Autorka w Twierdzeniu 7.2.2. w postaci dwóch nierówności, które powinna spełniać liczba losowanych z rozkładu równomiernego kolumn, obrabianych w Kroku I podejścia CSMC, oraz liczebność zbioru indeksów znanych elementów odtwarzanej macierzy. Oprócz rozmiarów tej macierzy, ważnym czynnikiem wchodzącym w sformułowanie warunków jest współczynnik (braku) koherencji odtwarzanej macierzy oraz wymagany przez nas poziom ufności, który jest tym bliższy 1 im bardziej współczynnik gamma jest większy od ln(3).

Dowód Twierdzeniu 7.2.2 jest bardzo rozbudowany. Doktorantka poprzedziła go szeregiem twierdzeń pomocniczych, z których najistotniejsze są Twierdzenia 7.2.4, inspirowane dowodem z pracy Xu [37] i Twierdzenie 7.2.6, które podaje dolne wymaganie na liczbę znanych elementów rekonstruowanej macierzy. Dowód tego twierdzenia jest bardzo interesujący, gdyż opiera się on na spostrzeżeniu, że hesjan kwadratu normy Frobeniusa odchylenia (funkcja g w rozprawie) daje się przedstawić jako iloczyn Kroneckera macierzy o mniejszych rozmiarach. Doktorantka wykazała się znakomitą znajomością własności iloczynów Kroneckera przy szacowaniu wartości własnych tego hesjanu.

W Rozdziale 8 mgr A. Krajewska przedstawia wyniki swoich badań swego podejścia i porównania go z innymi, dotąd znanymi algorytmami, na danych symulowanych. Jest to bardzo trafny wybór, gdyż, znając - na etapie generowania danych - dokładne wartości macierzy, można precyzyjnie ocenić dokładność ich rekonstrukcji.

Badania zostały starannie zaplanowane i obejmują bardzo szeroki zakres czynników, które mogą wpływać na dokładność rekonstrukcji i czas obliczeń. Należą do nich, między innymi, rozmiary macierzy, frakcja liczby znanych elementów macierzy, ułamek liczby kolumn losowanych w trakcie rekonstrukcji oraz szczegółowe wartości parametrów poszczególnych algorytmów rekonstrukcji. Dokładność rekonstrukcji oceniana była za pomocą błędu względnego norm macierzy oraz za pomocą ilorazu sygnału do szumu, który tutaj został odpowiednio zdefiniowany.

Przeprowadzone badania symulacyjne podzielić można na dwie grupy:

Grupa A) - to badania zaproponowanych algorytmów wykonane w takich warunkach, że spełnione są założenia twierdzeń z Rozdziału 7. Ich celem była ocena faktycznie osiąganej dokładności rekonstrukcji i czasu obliczeń, w zależności od wymienionych wyżej czynników eksperymentu. Ich realizacja wymagała, między innymi, opracowania algorytmu generowania macierzy zapewniających zadany stopień ich (nie-)koherencji.

Grupa B) – obejmuje badania porównawcze podejścia CSMC ze znanymi dotąd metodami uzupełniania macierzy, które uchodzą za odpowiednie do przybliżania macierzy o dużych rozmiarach.

Wyniki badań grupy A) stanowią przekonywującą dokumentację tezy rozprawy, iż przy spełnieniu założeń wymienionych w Rozdziale 7 możliwe jest uzyskiwanie gwarantowanej i bardzo wysokiej dokładności rekonstrukcji macierzy za pomocą podejścia CSMC, przy czym czas obliczeń jest istotnie krótszy niż dla metod opartych na dokładnej minimalizacji normy nuklearnej błędu za pomocą programowania półokreślonego.

Badania grupy B) realizowane były na macierzach 2000x3000, rzędu od 5 do 15 i ze stopniem znajomości elementów macierzy na poziomie 10-20%. Zarówno metody opracowane przez Doktorantkę algorytmy CSPGD-adam i CSPGD jak i algorytmy wybrane do porównań dawały błąd względny rzędu 0.001, z tym, że te oparte na podejściu CSMC dawały wyniki w czasie o około 160 sekund krótszym, przy zachowaniu tych samych warunków prowadzenia obliczeń.

W Rozdziale 9 mgr A. Krajewska raportuje wyniki zastosowania swoich algorytmów uzupełniania macierzy do problemów rzeczywistych:

- Budowy systemu rekomendacyjnego, w którym uzupełnianie macierzy służy jako narzędzie predykcji indywidualnych preferencji użytkownika, na podstawie preferencji znanych.

- Uzupełniania obrazów, których duża część elementów jest znacznie zniekształconych lub nieznanych.

- Predykcji nowych lub odtworzenia nieznanych połączeń w sieciach.

We wszystkich przypadkach konkluzje są podobne jak w Rozdziale 8, mimo że nie mamy gwarancji spełnienia wszystkich założeń teoretycznych. Najbardziej przemawiające do wyobraźni są wyniki uzupełniania obrazów rzeczywistych. W podrozdziale 9.2 Doktorantka dokumentuje znaczne, czasem nawet dwukrotne, skrócenie czasu obliczeń przy zastosowaniu swoich algorytmów w porównaniu z minimalizacją normy nuklearnej błędu

stosowanymi dotąd metodami. Co ciekawe, również wizualna ocena zrekonstruowanych obrazów wskazuje na zalety proponowanego w rozprawie podejścia.

Rozprawę kończy Rozdział 10, zawierający podsumowanie i kilka interesujących kierunków badań, które wykraczają poza jej zakres.

Ocena najważniejszych rezultatów rozprawy.

Cel postawiony przez mgr A. Krajewską został osiągnięty, a proponowane podejście CSMC i zaproponowane w oparciu o nie algorytmy są oryginalne i zostały zbadane pod kątem ich własności teoretycznych oraz zweryfikowane empirycznie na danych symulowanych i rzeczywistych. Realizacja celu rozprawy wymagała od Autorki zaproponowania szeregu oryginalnych rozwiązań szczegółowych oraz wiedzy i umiejętności z zakresu informatyki, algebry i metod numerycznych optymalizacji.
Osiągnięcia przedstawione w rozprawie mgr A. Krajewskiej oceniam bardzo wysoko. Proponuję rozważenie opublikowania rozprawy w formie książkowej w wydawnictwie o szerokim międzynarodowym obiegu.

Uwaga o charakterze dyskusyjnym

Nie mam uwag szczegółowych do treści rozprawy. Przedstawię tylko swoją uwagę o charakterze dyskusyjnym. Istotnym elementem dowodu w Rozdziale 7 jest spostrzeżenie, że hesjan funkcji celu da się przedstawić w postaci iloczynu Kroneckera macierzy o znacząco mniejszych rozmiarach. Czy istniałaby możliwość wykorzystania tego spostrzeżenia do dalszego przyspieszenia obliczeń metodą najmniejszych kwadratów ?

Pozostałe osiągnięcia Doktorantki

Mgr A. Krajewska jest autorką 2 artykułów w czasopismach o międzynarodowym obiegu i współautorką publikacji w materiałach międzynarodowej konferencji.

KONKLUZJA

Podsumowując całokształt rozprawy doktorskiej, oceniam ją jako bardzo znaczące osiągnięcie naukowe. W związku z tym stwierdzam, że rozprawa ta spełnia z nadmiarem wszystkie wymagania stawiane zwyczajowo i ustawowo rozprawom doktorskim, w szczególności, spełnia warunki i wymagania stawiane rozprawom doktorskim, określone w artykule 187 ust. 1 i ust. 2 Ustawy z dnia 20 lipca 2018 roku Prawo o Szkolnictwie Wyższym i Nauce (Dz.U. z 2018 poz. 1668 z

późn. zm.) i wnoszę o dopuszczenie jej do publicznej obrony.

Wysoką jakość wyników rozprawy oraz autorstwo publikacji w czasopismach o obiegu międzynarodowym powodują, że z pełnym przekonaniem wnioskuję o rozpatrzenie wniosku o wyróżnienie tej rozprawy.

Prof. Dr hab. inż. Ewaryst Rafajłowicz

Wrocław 14 września 2023 roku