Gdynia, 15 listopada 2024 r.

Prof. dr hab. inż. Ireneusz Czarnowski
Wydział Informatyki
Uniwersytet Morski w Gdyni
ul. Morska 83, 81-225 Gdynia

RECENZJA

rozprawy doktorskiej mgr. inż. Jacka KLIMASZEWSKIEGO

pt.: *„Uczenie modeli liniowych z regularyzacją dla małych zbiorów danych”.*

Recenzję przygotowałem w odpowiedzi na pismo Dziekana Wydziału Informatyki Zachodniopomorskiego Uniwersytetu Technologicznego w Szczecinie z dnia 25 września 2024 roku (nr pisma WI/Dokt-184/2024) informujące mnie o wyznaczeniu mojej osoby na recenzenta w przewodzie doktorskim mgr. inż. Jacka Klimaszewskiego.

1. Problematyka naukowa rozprawy

Rozprawa doktorska zatytułowana „*Uczenie modeli liniowych z regularyzacją dla małych zbiorów danych*” została przygotowana w dyscyplinie informatyka techniczna i telekomunikacja. Tematyka rozprawy dotyczy problemu nadzorowanego uczenia maszynowego. Szczególną uwagę Doktorant zwrócił na problem formowania liniowych modeli dyskryminacyjnych oraz regresyjnych z regularyzacją. Doktorant obszar badań ukierunkował także na aspekt radzenia sobie z sytuacją, gdy dane ucząca są relatywnie małe. Jest to również przypadek adekwatny do założenia, iż wykorzystując ich minimalny rozmiar nie tracimy na jakości predykcji. Praca rozważa także przypadki, gdy dane są małe w relacji do liczby atrybutów. Szukając odpowiedzi na stawiane pytania, Doktorant skupił się głównie na badaniu modelu regresji logistycznej, próbując przyśpieszyć procedurę uczenia. Zatem, praca ogranicza się również do aspektu strojenia modeli w kontekście zadania optymalizacyjnego.

Doktorant uzasadnia podjętą problematykę badania. Uzasadnia również aspekt regularyzacji w modelach liniowych, odwołując się przy tym do stosownej literatury. Formułuje cel oraz tezę pracy, która, uwzględniając parametry próby uczącej, takie jak rozmiar danych, rozkład danych czy korelacje, odwołuje się do możliwości opracowania algorytmów uczących wykorzystujących regularyzację i działających efektywniej w stosunku do istniejących rozwiązań.

Cel badawczy pracy jak i tezy, wokół których budowana jest struktura pracy doktorskiej, zostały sformułowane właściwie. Tematyka pracy jest aktualna. Należy również dodać, że problem poszukiwania efektywnych algorytmów uczenia maszynowego, ich strojenia oraz radzenia sobie z danymi o różnym charakterze i właściwościach jest ciągle otwartym problemem badawczym.

Wyniki badań Doktoranta zostały wcześniej zaprezentowane w kilku opracowaniach naukowych, dla których jest ich współautorem. Do opracowań tych należą następujące artykuły naukowe:

- Jacek Klimaszewski i Marcin Korzeń. “Fitting Penalized Logistic Regression Models Using QR Factorization”. W: Computational Science – ICCS 2020.Red.Valeria V. Krzhizhanovskaya i in. Cham: Springer International Publishing, 2020, s. 44–57. ISBN: 978-3-030-50417-5.
- Jacek Klimaszewski i Marcin Korzeń. “Image Smoothing Using ℓ^p Penalty for $0 \leq p \leq 1$ with Use of Alternating Minimization Algorithm”. W: Progress in Computer Recognition Systems. Red. Robert Burduk, Marek Kurzynski i Michał Wozniak. Cham: Springer International Publishing, 2020, s. 224–234. ISBN: 978-3-030-19738-4.
- Jacek Klimaszewski i Marcin Korzeń. “Optimization of ℓ^p regularized Linear Models via Coordinate Descent”. W: Schedae Informaticae 25 (2016), s. 61–72.
- Jacek Klimaszewski, Michał Sklyar i Marcin Korzeń. “Learning ℓ^1-penalized logistic regressions with smooth approximation”. W: 2017 IEEE International Conference on INnovations in Intelligent SysTems and Applications (INISTA). 2017, s. 126–130. DOI: 10.1109/INISTA.2017.8001144.
- Jan Rodziewicz-Bielewicz, Jacek Klimaszewski i Marcin Korzeń. “Regularized Learning of Neural Network with Application to Sparse PCA”. W: Artificial Intelligence and Soft Computing. Red. Leszek Rutkowski i in. Cham: Springer International Publishing, 2019, s. 203–214. ISBN: 978-3-030-20912-4.

2. Treść rozprawy

Rozprawa doktorska została przygotowana w języku polskim. Opracowanie łącznie składa się z 118 stron, w tym obejmuje:

- rozdział wprowadzający, w którym Doktorant przedstawił tło problemu oraz sformułował cel i tezę pracy,
- rozdział poświęcony uczeniu modeli liniowych,
- rozdział, w którym Doktorant odwołuje się do aspektu regularyzacji i proponuje swoje rozwiązania,
- rozdział nakreślający potencjalne obszary zastosowania proponowanego rozwiązania,
- rozdział prezentujący wyniki eksperymentu obliczeniowego,
- rozdział prezentujący zastosowanie regularyzacji w modelach liniowych,
- podsumowanie,
- spis literatury,
- załącznik A odwołujący się do kodów źródłowych zaimplementowanych metod.

Praca zawiera również streszczenie w języku polskim i angielskim oraz spis wybranych (podstawowych) symboli użytych w pracy.

Rozdział pierwszy rozprawy obejmuje sformułowanie problemu oraz wprowadza założenia dla dalszej dyskusji zawartej w pracy.

Rozdział drugi pracy Doktorant dedykował omówieniu aspektów uczenia modeli liniowych, prezentacji wybranych modeli liniowej dyskryminacji oraz regresji, idei regularyzacji oraz jej wariantów (unifikowanych funkcjami kary), oraz radzenia sobie z problemami szumu danych, przeuczenia modeli i braku jednoznacznej możliwości wskazania rozwiązania. Doktorant w rozdziale tym poddaje omówieniu różne metody uczenia modeli liniowych. Dokonuje analizy uczenia modelu regresji logistycznej z regularyzacją ℓ^2 za pomocą metody gradientowej oraz proponuje wprowadzenie rozkładu QR pozwalającego na redukcję rozwiązywanego problemu do określonej liczby próbek, co następnie obrazuje implementacją tego podejścia w jednej ze znanych bibliotek uczenia maszynowego. Następnie Doktorant analizuje możliwość zastosowania formuły Woodbury‘ego celem wyznaczenia rozkładu QR w przypadku macierzy rzadkich oraz redukcji kosztu obliczeniowego. Analizuje przy tym aspekt kosztów obliczeniowych tej możliwości. Szeroko omawia aspekt regularyzacji ℓ^1, odwołując się przy tym do optymalizacji funkcji kosztu, prezentuje metodę spadku względem

współrzędnych przeznaczoną do uczenia modelu regresji logistycznej z regularyzacją ℓ^1, podejście oparte na zastąpieniu normy ℓ^1 liniowymi ograniczeniami oraz podejście opierające się na zastąpieniu normy ℓ^1 jej górnym ograniczeniem za pomocą funkcji kwadratowej. W dalszej części rozdziału drugiego Doktorant omawia regularyzację zwaną „elastic net" oraz podejścia związane z wyznaczeniem parametrów funkcji kosztu dla tego wariantu regularyzacji. Kolejny podrozdział podejmuje temat aproksymacji normy ℓ^1. Doktorant uzasadnia podejście oraz charakteryzuje jego własności, a następnie dyskutuje sposób wyznaczania parametrów funkcji kosztu dla regularyzacji z użyciem pseudonorm ℓ^q, gdy $q \in [0,1]$. Rozważa wówczas analogicznie spadek względem współrzędnych oraz zastąpienie pseudonormy ℓ^q jej górnym ograniczeniem.

W rozdziale trzecim rozprawy Doktorant przedstawił ogólną koncepcję modelu sieci neuronowej oraz zaproponował model sieci neuronowej, o strukturze autoencodera, uczonej z regularyzacją ℓ^2. Następnie odniósł się do modelu uczenia ekstremalnego.

Czwarty rozdział pracy przedstawia uwarunkowania implementacyjne proponowanych rozwiązań z regularyzacją oraz ich walidację na drodze eksperymentów obliczeniowych. Innymi słowy rozdział ten zawiera wyniki eksperymentów oraz ich dyskusję. W pierwszym podrozdziale Doktorant przedstawia środowisko obliczeniowe przeprowadzenia eksperymentów, wykorzystane biblioteki oraz sposób dostosowania użytych procedur i funkcji dla potrzeb przeprowadzenia obliczeń. Następnie Doktorant przedstawił i omówił sposoby wyliczania składnika XX^T w przypadku macierzy rzadkich, co powiązał z modelem regresji logistycznej z regularyzacją ℓ^2 opartym na wykorzystaniu formuły Woodbury'ego. Doktorant przeanalizował dwa podejścia do wyliczenia tego składnika w przypadku macierzy rzadkich o podanych wymiarach oraz stopniu wypełnienia. Celem tego eksperymentu była ocena sposobu wyliczania wspomnianego iloczynu oraz wykorzystania potencjału proponowanego algorytmu uczenia modelu regresji logistycznej z regularyzacją ℓ^2. Następnie Doktorant odniósł się do oceny i wyboru sposobu implementacji metody gradientu sprzężonego, aby w kolejnym podrozdziale przedstawić przebieg eksperymentu dla modeli z regularyzacją ℓ^2 oraz ℓ^1. Celem tego eksperymentu była ocena wpływu liczby próbek *n* oraz wartości parametru regularyzującego λ (tj. siły kary w funkcji kosztu uczenia) na przebieg procesu ucznia. Doktorant dokonał oceny i identyfikacji czynników wpływających na czas działania algorytmów, ich złożoność oraz zbieżność. W konsekwencji, Doktorant przedstawia szereg wyników dla regularyzacji ℓ^2, wpływu parametru λ na przebieg procesu uczenia dla wybranych zestawów danych oraz różnych rozmiarów danych. Wyniki Doktorant przedstawia w formie tabelarycznej oraz na szeregu wykresów, na których porównuje czasy strojenia modeli, czasy wyznaczania ścieżek dla różnych λ oraz zbieżność procedur dla różnej liczby próbek. Procedurę oceny eksperymentalnej dla regularyzacji ℓ^2 doktorant powtórzył dla regularyzacji ℓ^1.

Rozdział piąty pracy, w pierwszym jego podrozdziale, dotyczy implementacji struktury rzadkiego encodera oraz jego oceny i porównania do metody PCA, jako metody referencyjnej. Eksperymenty przeprowadzono z wykorzystaniem wybranych danych benchmarkowych (danych analizy obrazowej), charakteryzujących się znacznie wyższą liczba atrybutów w porównaniu do wymiaru liczby przykładów (nazywanych przez Doktoranta liczbą próbek). Oceny rozwiązań Doktorant dokonał w odniesieniu do ich jakości lub czasu obliczeń, czy jakości rekonstrukcji obrazów, mając na uwadze przy tym poziom rzadkości rozwiązania. W oparciu o wyniki przeprowadzanych eksperymentów Doktorant zauważył, iż proponowane rozwiązanie uczenia struktury sieci neuronowej zachowuje wyższą dokładność rekonstrukcji dla większej rzadkości, a w niektórych przypadkach działa też szybciej. W drugim podrozdziale (podrozdział 5.2.) Doktorant eksperymenty oparł na modelu uczenia zrandomizowanego (uczenia ekstremalnego) oraz wykorzystał osiem modeli liniowych opartych na uczeniu z regularyzacją, porównując przy tym otrzymane wyniki z wynikami uzyskanymi przy użyciu modeli uczenia dostępnych w znanych bibliotekach uczenia maszynowego. Jednym z wniosków wypływających z tej części eksperymentów jest stwierdzenie o możliwości implementacji rozwiązania opartego na rozkładzie QR z uczeniem zrandomizowanym oraz z różnymi regularyzatorami , które może przyczynić się do poprawy działania klasycznych rozwiązań.

Ostatni z rozdziałów stanowi podsumowanie. Zawiera wnioski oraz dyskusję możliwości implementacji proponowanego podejścia z regularyzacją.

Literatura został dobrana właściwie dla podjętej problematyki, jest ona aktualna.

3. Wyniki uzyskane w pracy

Doktorant w swojej rozprawie doktorskiej skupił się na zagadnieniu uczenia modeli liniowych z regularyzacją oraz w oparciu o dane, dla których liczba atrybutów znacznie przewyższa liczbę obserwacji (przykładów, instancji, próbek). W badaniach skupił się głównie na modelu regresji logistycznej, próbując przyśpieszyć procedurę uczenia. Wnioskiem zasadniczym wypływającym z przeprowadzonego badania jest stwierdzenie, iż w zakładanych warunkach uczenia maszynowego regularyzacja jest jednym z głównych narzędzi poprawy własności numerycznych algorytmów, natomiast różne formy regularyzacji umożliwiają pozytywne kształtowanie własności obliczeniowych modeli wiedzy.

Do oryginalnych wyników uzyskanych przez Doktoranta zaliczyć należy:

- wykazanie, iż istnieje możliwość usprawnienia algorytmu uczenia dla modelu regresji z regularyzacją ℓ^2, poprzez zastosowanie rozkładu QR, a w sposób szczególny formuły Woodbury'ego,
- wykazanie możliwości wykorzystania rozkładu QR oraz formuły Woodbury'ego do uczenia modeli z rzadkimi funkcjami kary (typu lasso, elastic net),
- analizę wpływu ustawień i własności numerycznych algorytmów z uwagi na sposób reprezentacji zmiennopozycyjnej, metodę iteracyjnego rozwiązywania układów równań liniowych, sposób realizacji minimalizacji kierunkowej oraz sposobu reprezentacji macierzy danych (gęste, rzadkie),
- zaproponowanie algorytmu uczenia modelu regresji logistycznej z regularyzacją ℓ^2 opartego na spadku względem współrzędnych,
- zaproponowanie zastosowania technik uczenia modeli liniowych do strojenia modeli nieliniowych.

Należy dodać, że Doktorat w podsumowaniu pracy wskazuje również, na ograniczenia proponowanego rozwiązania opartego na regularyzacji oraz jej wariantach. Podkreśla również fakt, iż sformułowane wnioski zostały oparte jedynie na pewnej wybranej liczbie przykładów. Wskazuje również na pewne własności metod regularyzowanych, dających szersze możliwości na przykład z punktu widzenia interpretacji wyników. Należy również podkreślić, że prowadzona przez Doktoranta dyskusja wnosi szereg konkluzji dla potrzeb implementacji uczenia głębokiego, selekcji i ekstrakcji cech.

4. Pytania i uwagi do recenzowanej pracy

Recenzowana rozprawa doktorska podejmuje aktualny problem związany poszukiwaniem efektywnych algorytmów uczenia maszynowego, w tym wspomagających samo uczenie i konfigurowanie oraz radzenie sobie z problem danych uczących. Jest on aktualny również z punktu widzenia popularności głębokiego uczenia. Praca został przygotowana skrupulatnie oraz rozważa szereg aspektów. Niemniej jednak można sformułować kilka pytań i uwag, do których doktorant mógłby odnieść się podczas publicznej obrony.

Na ile zaproponowana metoda z regularyzacją jest odporna na minima lokalne, albo pozwala, poprzez jej zintegrowanie z innymi metodami, na radzenie sobie z tymi minimami?

W pracy wprost nie odniesiono się do aspektu rozwiązywania problemów wieloklasowych. Jedynie eksperymenty przeprowadzono na danych benchmarkach związanych z problemami zarówno binarnymi jak i wieloklasowymi. A zatem, czy kwestia rozwiązywania problemu dyskryminacyjnego jednoklasowego, binarnego lub wieloklasowego jest istotna w kontekście proponowanych implementacji uczenia z regularyzacją?

Na ile uzyskane wyniki, dla różnych porównywanych metod i ich wersji, różnią się istotnie statystycznie? Prezentując uzyskane wyniki, Doktorant pominął ich ocenę statystyczną, formułując wnioski jedynie na podstawie oceny uzyskanych wartości. Zatem zasadne jest ustosunkowanie się do pogłębionej analizy uzyskanych wyników.

Jak się ma kwestia zrównoleglenia obliczeń dla proponowanych metod regularyzowanych? Na ile metody te podatne są na ich wersje równoległe?

Doktorant mógłby podczas publicznej obrony odnieść się również do kierunków przyszłych badań powiązanych z uzyskanymi wynikami, które chciałby dalej kontynuować lub które w ogólności są do podjęcia.

5. Ocena redakcji i przygotowania rozprawy

Praca został przygotowane w sposób przejrzysty oraz poprawny językowo, chociaż w pracy występują nieliczne błędy stylistyczne i językowe. W pracy występują również powtórzenia (np., na stronach 13 i 16, czy 95 i 98).

Doktorant w pracy stosuje liczne oznaczenia i chociaż wprowadził spis oznaczeń, to spis ten mógł być nieco szerszy, aby ułatwić czytelnikowi poruszanie się po użytych oznaczeniach. W pracy występuję także mniej istotne niedociągnięcia, jak na przykład brak odwołania w treści pracy do Algorytmu 3.

Czytając pracę nasuwa się także konkluzja odnosząca się do braku wprowadzenia do rozdziału (gdy konkluzja ta dotyczy wszystkich rozdziałów), którego celem powinno być zapoznanie czytelnika z treścią danego rozdziału oraz przedstawienie kontekstu zawartego w danym rozdziale. Takie wprowadzenie pozwoliłoby na lepsze zrozumienie powiązania treści danego rozdziału z resztą pracy. Innym rozwiązaniem jest przedstawianie układu pracy we wprowadzeniu, czego również w recenzowanej pracy zabrakło.

Należy także dodać, że przedstawiając wyniki eksperymentów, a wpierw przedstawiając warunki przeprowadzenia tych eksperymentów, właściwą praktyką jest także sformułowanie celu i pytań badawczych. Takiego układu i opisu eksperymentu zabrakło w pracy, a mogłoby to dać lepsze zrozumienie przesłanek stojących za opisywanym eksperymentem obliczeniowym.

Pomimo tych uwag, redakcję pracy oceniam wysoko.

6. Konkluzja

Doktorant w swojej rozprawie doktorskiej podejmuje aktualny problem badawczy. Sformułowane przez Doktoranta wnioski, należy uznać za istotne z punktu widzenia rozwoju metod ucznia maszynowego oraz metod regularyzowanych.

Rozprawa prezentuje i potwierdza ogólną wiedzę teoretyczną odpowiednią dla osoby ubiegającej się o nadanie stopnia doktora. Uzyskane wyniki badania oraz wnioski zostały zaprezentowane w sposób właściwy dla oceny ich oryginalności i ważności dla dyscypliny informatyka techniczna i telekomunikacja. Stwierdzam, że Doktorant wykazał się umiejętnością samodzielnego rozwiązania problemu badawczego.

Podsumowując, uważam, że rozprawa doktorska mgr. inż. Jacka Klimaszewskiego pt. ”*Uczenie modeli liniowych z regularyzacją dla małych zbiorów danych*” spełnia wymogi stawiane przy ubieganiu się o nadanie stopnia doktora w dyscyplinie informatyka techniczna i telekomunikacja. Tym samym wnioskuję o dopuszczenie rozprawy doktorskiej do publicznej obrony.

Ireneusz Czarnowski

Dokument podpisany przez Ireneusz Czarnowski
Data: 2024.11.17 22:18:15 CET