Prof. dr hab. inż. Andrzej Skowron
Instytut Badań Systemowych PAN

Warszawa, 05. 12. 2023

Recenzja rozprawy doktorskiej mgr. inż. Mikołaja Markiewicza pt. ”Evaluation of data partitioning strategies for distributed clustering and classification algorithms”

Recenzja niniejsza została sporządzona w odpowiedzi na pismo Przewodniczącego Rady Naukowej Dyscypliny Informatyka Techniczna i Telekomunikacja Politechniki Warszawskiej Prof. dr. hab. inż. Jarosława Arabasa z dnia 21 września 2023 roku zawierającego prośbę o zrecenzowanie rozprawy doktorskiej mgr. inż. Mikołaja Markiewicza.

Rozprawa jest zatytułowana ”Evaluation of data partitioning strategies for distributed clustering and classification algorithms” i została przygotowana pod kierunkiem dr. hab. inż. Piotra Gawrysiaka z udziałem promotora pomocniczego dr. Jakuba Koperwasa. Została ona opublikowana w roku 2023 przez Wydawnictwo Politechniki Warszawskiej. Rozprawa jest napisana w języku angielskim i liczy 119 stron. Jest ona podzielona na 6 rozdziałów i zawiera czteroczęściowy załącznik a także spis treści, obszerny spis literatury oraz streszczenia w języku polskim i angielskim.

1. Uwagi ogólne nt. problematyki rozprawy i jej celów

Celem rozprawy jest opracowanie nowych metod mających na celu poprawę ewaluacji algorytmów rozproszonych analizy danych poprzez zastosowanie nowych metod nierównomiernego (non-IID) partycjonowania danych w celu umożliwienia badania wpływu niezdefiniowanego rozkładu danych na optymalizację przetwarzania.

Problematyka rozprawy obejmuje bardzo aktualne zagadnienia występujące, np. w uczeniu federacyjnym (ang. federated learning). Analizowane w rozprawie problemy rozproszonego przetwarzania danych przy braku informacji o ich rozkładach stanowią spore wyzwanie. Tematyka ta może niewątpliwie stanowić przedmiot badań rozprawy doktorskiej.

2. Ocena formy i struktury rozprawy

Rozdział 1 stanowi wstęp do rozprawy i przedstawia motywację, cel rozprawy oraz omówienie wyników, ograniczeń przyjętych dla podejścia analizowanego w rozprawie oraz krótkie omówienie

treści. Rozdział 2 zawiera podstawowe pojęcia dla rozprawy, w szczególności dotyczące podziału danych i podobieństwa, typów rozważanych algorytmów oraz rozproszonego przetwarzania danych. Podsumowanie aktualnych prac nad benchmarkami dla algorytmów rozproszonych przynosi Rozdział 3. Główne wyniki rozprawy zawarte są w Rozdziałach 4 i 5. Rozdział 4 przedstawia opracowaną w rozprawie metodę ewaluacji i algorytm, w szczególności nowe metody podziału danych, hybrydowy algorytm klastrowania oraz omówienie opracowanej w rozprawie platformy. Rozdział 5 jest raportem z przeprowadzonych eksperymentów bazującym na zaproponowanym w rozprawie podejściu do ich ewaluacji. Rozdział 6 stanowi podsumowanie i przedstawia kierunki dalszych badań. Dodatki zawarte w Rozdziale 7 zawierają szczegóły techniczne dotyczące DDM-PS-Eval, uwagi o komunikacji, konfiguracji środowiska, szczegóły odnośnie jakości wyników oraz statystyk odnośnie czasu przetwarzania oraz transferu obciążenia.

3. Osiągnięcia badawcze doktoranta

Doktorant zrealizował postawiony w rozprawie cel. Rozprawa jest napisana poprawnie choć niespecjaliści z pewnością będą mieli problem ze zrozumieniem niektórych jej fragmentów.

Do głównych wyników rozprawy zaliczam:

- Opracowanie podejścia ewaluacyjnego bazującego na algorytmach partycjonowania danych wraz z eksperymentalną analizą ich wpływu na jakość algorytmu realizowanego w środowisku rozproszonym; w szczególności opracowanie metod symulacji niejednorodnego rozpraszania zbiorów danych na niezależne jednostki obliczeniowe (szczególnie uwzględniające wymienione w pracy [33] ze spisu prac w rozprawie takie aspekty jak: feature distribution skew (covariate shift), label distribution skew (prior probability shift), same label, different features (concept drift), same features, different label (concept shift), quantity skew (unbalancedness)), co pozwala na estymację poprawności algorytmu rozproszonego bez założenia o jednorodnym rozkładzie zbiorów danych pomiędzy jednostki systemu rozproszonego.

- Opracowanie hybrydowego algorytmu klastrowania (analizy skupień) zbiorów danych dostosowanego do pracy w środowisku rozproszonym danych z niejednorodnymi rozkładami. Jako bazowy używany jest rozproszony algorytm k-średnich a algorytm OPTICS zastosowany jest do konstrukcji globalnych modeli klastrowania. Globalne klastrowanie danych uzyskuje się z lokalnych modeli na podstawie lokalnych statystyk z wykorzystaniem najlepszych podziałów danych i podejścia bazującego na gęstościach. Algorytm nie wymaga załadowania

wszystkich danych do jednostki centralnej a koszt komunikacji jednostek obliczeniowych jest znikomy. Przeprowadzone eksperymenty wykazały pozytywne cechy opracowanego algorytmu w porównaniu z istniejącymi rozwiązaniami do których należą szybkość działania, niezawodność i wysoka jakość przy braku założenia o rozkładnie danych. Eksperymenty z przedstawionym w rozprawie algorytmem hybrydowego klastrowania wskazały na możliwość uzyskania przetwarzania o wysokiej jakości i rozwiązania problemu partycjonowania określonych danych, bez konieczności przyjmowania założeń dotyczących jednorodnej dystrybucji danych pomiędzy jednostki systemu rozproszonego.

- Opracowanie ewaluacyjnej platformy dla rozproszonych algorytmów eksploracji danych (ang. data mining) zapewniającej analizę wyników klastrowania i klasyfikacji, z uwzględnieniem różnych wymienionych wyżej aspektów rozpraszania zbiorów danych, z uwagi na finalną jakość, obciążenie transferu i szczegółowe pomiary czasu w poszczególnych etapach przetwarzania. Rozprawa zawiera stosunkowo obszerny raport z przeprowadzonych eksperymentów wraz z dyskusją uzyskanych wyników wskazująca na zalety opracowanego podejścia. Opracowana platforma, dedykowana do ewaluacji rozproszonych algorytmów z non-IID rozkładami danych, jest jak autor zapewnia łatwą w obsłudze, rozbudowywaną platformę umożliwiającą dynamiczne dołączania komponentów i rozwiązywanie problemów co jest istotne wobec braku kompleksowych narzędzi dla analizy porównawczej metod eksploracji danych rozproszonych (DDM).

Warto dodać, że autor rozprawy opublikował już kilka prac związanych z tematyką rozprawy.

4. Uwagi krytyczne i dyskusyjne

Wymienię tu kilka uwag krytycznych i dyskusyjnych.

- Autor rozprawy nie zadbał dostatecznie o zapewnienie reprodukowalności wyników (ang. reproducibility) poprzez udostępnienie platformy i zbiorów danych dla powtórzenia eksperymentów przez użytkowników platformy.
- Zabrakło mi w rozprawie odniesienia do metody statystycznej ewaluacji różnic uzyskiwanych w eksperymentach z zastosowaniem różnych metod.
- Pewien niedosyt związany jest z brakiem głębszej dyskusji o stosowanych miarach jakości klasyfikatorów czy klastrów, np. w przypadku danych niezbalansowanych.

- Zabrakło mi głębszej dyskusji o tym czy i jak zaproponowane podejście pozwala oceniać ryzyko, że wprowadzone niejednorodne dane zniekształcają w tak istotny sposób uzyskane wyniki, że czynią przyjęte rozwiązanie zawodnym.
- Brak szerszej dyskusji na ile wymienione wyżej aspekty niejednorodności danych wyczerpują sytuacje w rzeczywistości.
- Jak można odnieść uzyskane wyniki do sytuacji rzeczywistej gdy nie mamy do dyspozycji globalnych zbiorów danych?
- Ostatnia uwaga odnosi się do znanej pracy pt. ”The mathematics of learning. Dealing with data” autorstwa Tomaso Poggio and Steven Smale (Notices of the American Mathematical Society (AMS), Vol. 50, No. 5, 537-544, 2003), w której autorzy, m.in. wskazują, że ‘jednopoziomowa’ (jednorazowa) agregacja lokalnych modeli może zawodzić w przypadku złożonych problemów (np. rozpoznawania postaci na drodze) i bardziej właściwe byłoby opracowanie modeli hierarchicznych na wzór mózgu człowieka (niestety ta struktura nie jest jeszcze dostatecznie poznana). Czy w przypadku problemów analizowanych w rozprawie podejście hierarchiczne (nawet na niewielkie głębokości) nie prowadziłoby do poprawy wyników (nasuwa się tu, np. analogia dotyczącą analizy języka naturalnego z zastosowaniem n-gramów)? To pytanie wiąże się też z rolą wnioskowania aproksymacyjnego, wzdłuż którego mogłyby być generowane pewne istotne lokalne agregacje modeli lokalnych (z ograniczeniem do pewnej niewielkiej głębokości) i prowadzić stopniowo do poprawy wyników.

5. Konkluzja końcowa

Powyższe uwagi krytyczne nie wpływają na moją pozytywną opinię o rozprawie, niektóre z nich mają charakter dyskusyjny. Cel rozprawy został zrealizowany.

W moje ocenie rozprawa doktorska mgr. inż. Mikołaja Markiewicza pt. ”Evaluation of data partitioning strategies for distributed clustering and classification algorithms” spełnia warunki stawiane przez odpowiednią ustawę w odniesieniu do rozpraw doktorskich i może być dopuszczona do publicznej obrony.