RECENZJA ROZPRAWY DOKTORSKIEJ

I. Dane identyfikacyjne rozprawy

Tytuł rozprawy: Classification based on dispersed data with deep learning issues

Autor: Kwabena Frimpong Marfo

Jednostka / instytut: Uniwersytet Śląski w Katowicach, Instytut Informatyki

Promotor: dr hab. Małgorzata Przybyła-Kasperek, prof. UŚ

Rok: 2025

Forma: Rozprawa kumulatywna (na podstawie publikacji; 11 prac, lata 2021–2025)

II. Przedmiot, cele i zakres pracy

Rozprawa dotyczy problemu klasyfikacji nadzorowanej w warunkach, w których dane są rozproszone pomiędzy niezależnymi źródłami (silosami) i nie mogą zostać scentralizowane ze względu na ograniczenia prawne, organizacyjne i/lub prywatności. Szczególny nacisk położono na sytuacje heterogeniczności strukturalnej: różne zbiory obiektów oraz różne zbiory atrybutów warunkowych w poszczególnych źródłach, przy wspólnym atrybucie decyzyjnym.

Celem głównym jest zaproponowanie i zweryfikowanie metod uczenia modeli klasyfikacyjnych w takich warunkach w sposób ograniczający wymianę informacji, w szczególności bez przekazywania surowych danych pomiędzy podmiotami.

III. Sformułowanie problemu badawczego i pytania badawcze

Autor formułuje hipotezę, że odpowiednio zaprojektowane architektury oparte na sieciach neuronowych mogą istotnie poprawić skuteczność klasyfikacji dla danych rozproszonych i heterogenicznych, w porównaniu z podejściami klasycznymi oraz z typowymi wariantami uczenia federacyjnego zakładającymi zgodność przestrzeni cech.

Rozprawa organizuje wkład naukowy w trzy linie metodologiczne odpowiadające trzem pytaniom badawczym:

A. fuzja predykcji lokalnych (wektorów prawdopodobieństw klas) z użyciem sieci MLP oraz RBF;
B. agregacja parametrów lokalnych sieci neuronowych po ujednoliceniu przestrzeni wejściowej przez imputację brakujących atrybutów z wykorzystaniem sztucznie generowanych obiektów;

C. hierarchiczny schemat: lokalna ekstrakcja cech (PCA/SVD/UMAP) do wymiaru k, uczenie lokalnych modeli oraz agregacja na poziomie decyzji (soft voting).

IV. Omówienie zastosowanych metod i rozwiązań

W ramach podejścia A (prace P1–P4, P7, P10) każde źródło danych buduje lokalny klasyfikator k-NN (z miarą podobieństwa Gowera dla cech mieszanych), generując wektory prawdopodobieństw przynależności do klas. Wektory te są następnie łączone i poddawane fuzji przez sieć MLP lub RBF; w wariancie RBF zastosowano inicjalizację centrów metodą k-średnich oraz heurystykę doboru szerokości jąder, a dla danych obrazowych rozszerzono podejście o nadzorowane uczenie centrów.

W metodzie B (P5, P6, P9) zaproponowano mechanizm imputacji brakujących atrybutów dla danych heterogenicznych poprzez konstrukcję sztucznych obiektów na podstawie statystyk warunkowanych klasą (np. MIN/MAX/AVG/MED). Po ujednoliceniu wejść trenowane są lokalne sieci MLP o zgodnej architekturze, a następnie agregowane do postaci modelu globalnego poprzez łączenie parametrów.

W podejściu C (P8, P11) w każdym źródle wykonywana jest lokalna projekcja danych do wspólnego wymiaru k (PCA, SVD lub UMAP), następnie trenowany jest lokalny model (m.in. MLP, SIMPLE, GRU, LSTM, RBF). Wnioskowanie odbywa się przez agregację predykcji z wielu źródeł metodą soft voting, bez konieczności wyrównywania współrzędnych przestrzeni cech pomiędzy silosami.

V. Ocena oryginalności i znaczenia pracy

Praca wnosi oryginalny wkład poprzez zaproponowanie oraz szeroką weryfikację empiryczną kilku komplementarnych strategii uczenia w warunkach rozproszenia i heterogeniczności. Na szczególne podkreślenie zasługują: fuzja predykcji lokalnych z użyciem sieci RBF i jej wariantu z uczeniem centrów dla danych obrazowych, ujednolicenie przestrzeni wejściowej poprzez sztucznie generowane obiekty umożliwiające agregację parametrów sieci, oraz hierarchiczny schemat lokalnej ekstrakcji cech połączony z agregacją decyzji.

Zagadnienie ma wysoką aktualność aplikacyjną w obszarach, gdzie centralizacja danych jest utrudniona (ochrona zdrowia, finanse, administracja publiczna), a jednocześnie istnieje potrzeba budowy modeli o dobrej jakości predykcyjnej.

VI. Ocena poprawności metodycznej, eksperymentów i wniosków

Rozprawa przedstawia formalizację danych rozproszonych oraz adekwatnie różnicuje omawiany problem względem klasycznego uczenia rozproszonego i typowych schematów federacyjnych. Autor przeprowadza rozbudowane badania eksperymentalne na zbiorach benchmarkowych (w tym UCI) oraz na danych syntetycznych, analizując m.in. poziom rozproszenia, szum i niezrównoważenie klas. Zastosowano właściwe testy statystyczne nieparametryczne (np. Friedman/Wilcoxon/Kruskal–Wallis) do oceny istotności różnic.

Wnioski w większości znajdują potwierdzenie w przedstawionych wynikach: warianty oparte na sieci RBF w podejściu A często przewyższają MLP przy mniejszej złożoności; metoda B pokazuje, że imputacja i agregacja parametrów mogą działać skutecznie przy odpowiednim doborze liczby sztucznych obiektów; metoda C demonstruje konkurencyjność podejścia z lokalną redukcją wymiaru i głosowaniem miękkim. Jednocześnie część obserwacji ma charakter zależny od danych i wymaga ostrożnej generalizacji.

VII. Uwagi krytyczne

1. Sformułowanie „ujednolicona przestrzeń cech" w linii C może sugerować wyrównanie współrzędnych pomiędzy silosami. Ponieważ transformacje PCA/UMAP są liczone lokalnie, przestrzenie te nie są współrzędnie zgodne; w praktyce mamy agregację na poziomie decyzji. Zalecane jest doprecyzowanie terminologii oraz rozważenie eksperymentu z wariantem, w którym elementy transformacji są współdzielone w minimalnym zakresie.

2. Aspekt prywatności ma charakter pragmatyczny (brak wymiany surowych danych), lecz nie obejmuje formalnych gwarancji (np. DP, secure aggregation, MPC/HE). Warto wyraźnie oddzielić „prywatność organizacyjną" od „prywatności formalnej" oraz wskazać potencjalne kierunki rozszerzeń.

3. Skalowalność i weryfikacja na danych rzeczywistych o dużej skali: eksperymenty oparto głównie na zbiorach małych/średnich. Dla wzmocnienia wiarygodności wdrożeniowej przydałaby się analiza kosztów obliczeniowych i komunikacyjnych oraz testy na większych, wysokowymiarowych danych.

4. Metoda B: imputacja na podstawie statystyk warunkowanych klasą może prowadzić do wygładzania rzadkich wzorców (np. klas mniejszościowych) i potencjalnie ujawniać informacje o rozkładach. Warto zastanowić się nad ryzykami oraz porównać z alternatywami (np. imputacja modelowa, modele generatywne).

IX. Pytania do Autora na publiczną obronę

1. W metodzie C: na ile uzasadnione jest określenie „ujednolicona przestrzeń cech" przy lokalnych transformacjach PCA/UMAP? Czy częściowe współdzielenie parametrów transformacji (np. komponentów) poprawiłoby wyniki bez istotnego naruszenia ograniczeń?

2. Jakie ryzyka prywatności wiążą się z przekazywaniem statystyk warunkowanych klasą w metodzie B i jak można je ograniczyć (np. poprzez agregację bezpieczną lub mechanizmy DP)?

5. Jakie byłyby potencjalne korzyści i ryzyka zastosowania modeli generatywnych (VAE/GAN) do imputacji w linii B w porównaniu z podejściem deterministycznym?

X. Konkluzja i wniosek recenzenta

Rozprawa stanowi oryginalny, dojrzały i wartościowy wkład w obszar uczenia maszynowego z danych rozproszonych i heterogenicznych. Na szczególne podkreślenie zasługuje wysoki poziom merytoryczny pracy, poprawność metodyczna oraz konsekwentnie przeprowadzona weryfikacja empiryczna.

Dorobek publikacyjny Autora należy ocenić bardzo wysoko — jest on nie tylko obszerny, ale również spójny tematycznie, dobrze zaplanowany i świadczy o samodzielności badawczej oraz umiejętności prowadzenia badań w sposób systematyczny. W mojej ocenie prezentowane wyniki i aktywność publikacyjna wskazują na ponadprzeciętną dojrzałość naukową doktoranta.

Konkluzja: Praca spełnia standardy wymagane do uzyskania stopnia doktora w dyscyplinie informatyka.

Rekomendacja: Przyjąć rozprawę i dopuścić Autora do publicznej obrony.

Recenzent: dr hab. inż. Rafał Deja, prof. AWSB

Miejscowość i data: Katowice, 22.04.2026