# Streszczenie rozprawy doktorskiej

Badania nad środowiskami danych rozproszonych koncentrują się na opracowywaniu modeli klasyfikacji zdolnych do działania w sytuacjach, gdy dane są rozproszone pomiędzy wieloma niezależnymi źródłami, często o heterogenicznych strukturach, których nie można scentralizować ze względu na przepisy dotyczące ochrony danych, takie jak RODO. Zbiory danych spotykane w rzeczywistych zastosowaniach rzadko są jednorodne — charakteryzuje je zróżnicowanie zestawów cech, odmienne rozkłady obiektów oraz ścisłe ograniczenia prywatności. Klasyczne metody uczenia maszynowego, oparte na założeniu jednorodnych i centralnie dostępnych danych, okazują się niewystarczające w takich warunkach. Wobec tego opracowanie procedur uczenia odpornych na strukturalną niekompatybilność i rozproszenie danych staje się warunkiem koniecznym do wykorzystania takich fragmentarycznych zbiorów w celu zbudowania dokładnych i użytecznych systemów klasyfikacji.

W rozprawie zaproponowano sekwencję komplementarnych metod, które adresują różne aspekty uczenia w warunkach rozproszenia. Każde z opracowanych podejść wynikało bezpośrednio z ograniczeń zaobserwowanych w rzeczywistych systemach, gdzie danych nie można było udostępniać ani ujednolicić w prosty sposób. Opracowane algorytmy przebadano zarówno na zbiorach syntetycznych, jak i rzeczywistych, wykazując ich zgodność z założeniami oraz zdolność do zapewnienia dokładnej i stabilnej klasyfikacji nawet w silnie heterogenicznych środowiskach.

Pierwsza grupa badań koncentrowała się na fuzji predykcji lokalnych. Gdy każde niezależne źródło danych trenuje własny klasyfikator, wyzwaniem jest uzyskanie spójnej globalnej predykcji bez wymiany surowych danych. W tym celu zaproponowano metodę, w której na podstawie każdego lokalnego zbioru danych trenowany jest klasyfikator k-NN, a otrzymane wektory predykcji są następnie integrowane przy użyciu sieci neuronowej. Podejście to nie tylko redukowało problemy związane z prywatnością i heterogenicznością strukturalną danych, ale dzięki komponentowi sieci neuronowej skutecznie wychwytywało złożone wzorce, generując jednoznaczne decyzje. Metodę tę rozszerzono, pokazując wpływ intensywności szumu, zróżnicowanych charakterystyk danych oraz różnych poziomów rozproszenia na stabilność procesu fuzji, potwierdzając odporność podejścia na zakłócenia i zróżnicowanie stopnia rozproszenia. Następnie ulepszono tę koncepcję, zastępując tradycyjną fuzję MLP sieciami RBF, wraz z dedykowaną strategią strojenia centrów RBF. Znacząco zmniejszyło to złożoność architektury, przyspieszyło zbieżność i ograniczyło ryzyko przeuczenia, które jest typowe dla metod opartych na MLP. Te rezultaty dostarczyły zestawu niezawodnych narzędzi do łączenia predykcji pochodzących z heterogenicznych lokalnych zbiorów danych, nawet gdy ich struktury znacząco się różnią.

Druga grupa publikacji przesunęła punkt ciężkości w stronę budowy pojedynczego globalnego modelu z rozproszonych zbiorów danych. Zamiast polegać na fuzji predykcji, celem było ujednolicenie struktury lokalnych tablic decyzyjnych tak, aby możliwa była agregacja lokalnych sieci neuronowych w jedną globalną sieć. Kluczową trudnością było to, że lokalne tablice mogą zawierać całkowicie odmienne

zestawy atrybutów warunkowych. W celu rozwiązania tego problemu zaproponowano metodę imputacji brakujących atrybutów opartą na sztucznie generowanych obiektach, tworzonych na podstawie statystycznych właściwości pozostałych lokalnych tablic. Strategia ta umożliwia trenowanie lokalnych sieci neuronowych o zgodnych strukturach, które następnie można połączyć w jednym kroku agregacji parametrów. Dalsze rozszerzenie tej koncepcji obejmowało analizę liczby i zróżnicowania sztucznych obiektów oraz opracowanie wytycznych dotyczących projektowania skutecznych architektur MLP wykorzystywanych przy konstrukcji globalnych modeli.

Ostatnia grupa badań odchodziła od imputacji obiektów w tablicach lokalnych na rzecz harmonizacji opartej na ekstrakcji cech. Zamiast rozszerzać zbiory danych o obiekty syntetyczne, każdą lokalną tabelę przekształcano we wspólną przestrzeń cech przy użyciu metod redukcji wymiarów. W szczególności zaproponowano zastosowanie PCA do odwzorowania każdego zbioru danych w spójną, k-wymiarową reprezentację, co umożliwia późniejszą fuzję predykcji modeli MLP poprzez głosowanie przy jednoczesnym zachowaniu prywatności danych. Koncepcję tę znacznie rozszerzono, obejmując szerszy zestaw metod ekstrakcji cech, takich jak PCA, SVD i UMAP, oraz oceniając systematycznie różne architektury sieci neuronowych, w tym MLP, RBF, GRU, LSTM i SIMPLE. Badania te wykazały, że harmonizacja oparta na cechach zapewnia stabilne i dokładne wyniki klasyfikacji, szczególnie gdy wybrana reprezentacja zachowuje bogaty zbiór cech informatywnych.

Obecnie duże, rozproszone systemy informacyjne odgrywają kluczową rolę w przetwarzaniu danych w wielu dziedzinach. Jednak klasyczne techniki uczenia maszynowego często nie są dostosowane do rzeczywistych środowisk charakteryzujących się niejednolitością cech, ścisłymi wymogami prywatności i fragmentacją danych. Przedstawione w rozprawie osiągnięcia oferują zestaw nowych podejść, które przełamują te bariery. Proponując metody fuzji predykcji, agregacji parametrów oraz unifikacji przestrzeni cech i weryfikując je w szerokim zakresie scenariuszy rozproszonych, badania te poszerzają praktyczne zastosowania sieci neuronowych w niejednorodnych, wrażliwych na prywatność ustawieniach. Opracowane rozwiązania otwierają nowe kierunki dla zdecentralizowanego uczenia i pokazują, że dokładne modele globalne można budować bez naruszania lokalności danych, znacząco zwiększając użyteczność technik uczenia maszynowego we współczesnych, rozproszonych systemach.