Abstract

This thesis presents a series of publications that addresses the challenge of constructing optimal global classification models from dispersed and heterogeneous datasets without requiring centralized data access. Here, dispersed or distributed data refers to datasets stored across multiple independent sources that cannot be physically consolidated, while heterogeneous data denotes collections whose feature spaces and object distributions differ across locations. Traditional approaches to dispersed data environments assume uniform feature spaces and direct data aggregation, making them unsuitable for real-world scenarios involving privacy constraints and structural data variability. To overcome these limitations, this research introduces and evaluates three complementary neural network approaches, each designed to address the challenges of non-homogeneous and distributed data. The first approach integrates predictions from locally trained k-nearest neighbors classifiers through a neural network that performs a fusion of independent outputs, achieving a unified global classification decision while preserving data privacy. The second approach employs an artificial-object imputation mechanism that harmonizes structural differences among local datasets, enabling aggregation of the trained network weights into a single global model. Lastly, the third approach applies feature-extraction techniques such as Principal Component Analysis (PCA), Singular Value Decomposition (SVD) and Uniform Manifold Approximation and Projection (UMAP) to project local datasets into a unified feature space, after which the resulting models are combined using a soft voting scheme. Experimental evaluations demonstrate that these approaches consistently enhance classification accuracy and robustness compared to classical ensemble and centralized methods, while ensuring privacy preservation and computational efficiency. The findings contribute novel insights into neural network-based distributed learning and establishes a foundation for scalable, privacy-aware data integration in dispersed data environments.