Doctoral Dissertation Summary

Research on dispersed data environments seeks to develop classification models capable of operating when data are distributed across multiple independent sources, often with heterogeneous structures that cannot be centralized due to data privacy laws such as GDPR. Real-world datasets are rarely homogeneous, characterized by varying feature sets, varying object distributions and strict privacy constraints. Classical machine learning methods, which typically assume homogeneous and centrally available data, struggle under such conditions. Consequently, designing learning procedures that are robust to structural incompatibility and dispersion is essential for transforming these fragmented datasets into accurate and practical classification systems.

In this thesis, a sequence of complementary methods addressing different facets of dispersed learning was proposed. Each of the developed approaches emerged directly from limitations observed in real systems, where data could not be shared or aligned in a trivial way. The resulting algorithms were examined using both synthetic and real datasets and were shown to behave as intended, delivering improved and consistent classification performance even in highly heterogeneous settings.

The first group of studies focused on the fusion of local predictions. When each independent data source train its own classifier, the challenge lies in generating a coherent global decision without exchanging raw attributes. To address this, a method was introduced in which each local dataset trains a k-NN classifier, and the resulting prediction vectors are subsequently integrated using a neural network. This approach not only mitigated the challenges related to data privacy and structural heterogeneity of data, but with the neural network component effectively capturing intricate patterns, the approach generated unambiguous decisions. The method discussed above was extended to demonstrate how noise intensities, varying data characteristics, and different levels of dispersion affect the stability of the fusion process, showcasing the robustness of the approach to noise perturbations and various data dispersion levels. Later, this approach was enhanced by replacing traditional MLP fusion with RBF networks, including a dedicated training strategy for tuning RBF centers. This substantially reduced architectural complexity, facilitated faster convergence, and mitigated the risk of overfitting commonly associated with MLP-based methods. These contributions collectively provided a set of reliable tools for fusing predictions originating from heterogeneous local datasets, even when the underlying data structures differ substantially.

The next group of publications shifted the focus toward constructing a single global model from dispersed datasets. Instead of relying on prediction fusion, the objective was to standardize the structure of local decision tables so that local neural networks could be aggregated into a global one. A key difficulty addressed here stems from the fact that local tables may contain entirely different sets of conditional attributes. To solve this, a constructive imputation method based on artificially generated objects that fill missing attributes using statistical characteristics gathered from other local tables. This strategy makes it possible to train structurally aligned neural networks locally and later combine their parameters in a single aggregation step. These ideas were expanded, where the quantity and diversity of artificial objects were analyzed in depth, and guidelines for designing effective MLP architectures for such global model construction were proposed.

The last group of studies shifted from artificial imputation of local tables to feature-extraction based harmonization. Instead of augmenting datasets with synthetic objects, each local table was transformed into a shared feature space using dimensionality reduction methods. In particular, the use of PCA was proposed to map every dataset into a consistent k-dimensional representation, allowing subsequent fusion of predictions from MLP models through soft voting, while ensuring data privacy. This idea was significantly expanded to accommodate a broader suite of feature-extraction methods including PCA, SVD and UMAP and neural architectures such as MLP, RBF, GRU, LSTM, SIMPLE were systematically evaluated. These studies demonstrated that feature-based harmonization produces stable, high-quality classification results, particularly when the chosen representation retains a rich set of informative features.

Today, large-scale, distributed information systems are central to data processing in many domains. However, classical machine-learning techniques often remain ill-suited for real-world environments characterized by feature inconsistency, strict privacy requirements and fragmented data. The contributions presented in this thesis collectively offer a set of novel approaches that overcome these barriers. By proposing methods for prediction fusion, parameter aggregation and feature-space unification, and by validating them across a wide range of dispersed scenarios, the research expands the practical applicability of neural networks to non-uniform, privacy-sensitive settings. These advancements open new directions for decentralized learning and demonstrate that accurate global models can be constructed without violating data locality, thus significantly broadening the usability of machine-learning techniques in modern distributed systems.