EDITORIAL

Non-IID Learning

Longbing Cao, *Editor in Chief*

Real-life AI systems are non-IID, i.e., their variables are unlikely independent and drawn from the same distribution. Instead, non-IIDness is a common characteristic and complexity of real-life systems, where variables, objects, and subsystems are coupled/interactive and heterogeneous. This issue highlights this important theme on Non-IID Learning with six feature articles. In addition, four columns highlight expert opinions on beyond i.i.d., trustworthy AI, data-driven predictive maintenance, and secrets for data science deployments, respectively.

In this issue, I am pleased to present four columns: Editor's Perspective, AI Expert, AI Focus, and AI Insight, respectively. The theme spotlights six articles on an important but often overlooked or simplified topic—Non-IID Learning—in AI, analytics, learning, and more broadly, almost every science and engineering discipline.

COLUMN AND DEPARTMENT ARTICLES

In the Editor's Perspective column, I discuss the topic of "Beyond i.i.d.: Non-IID Thinking, Informatics, and Learning." Almost all science, technology, engineering, and their applications commonly follow the IID assumption, which simplifies the often complicated reality and complexities in real life. The resultant theories, techniques, and systems may not support or approach reality-level, actionable problem-solving. Here, I discuss the IID to non-IID paradigm shift, and the general frameworks of IID thinking and non-IID thinking, and propose general conceptual and research maps of non-IIDness and non-IID learning. The challenges, issues, and opportunities in quantifying interactions, heterogeneities, complexities, and intelligence are discussed. Such non-IID thinking and developments are essential for non-IID statistics, informatics, and computing.

AI Expert presents the article "From Features Engineering to Scenarios Engineering for Trustworthy AI: I&I, C&C, and V&V," by the column editor Fei-Yue Wang and his coauthors. The article introduces a theoretical framework of *scenarios engineering* with six dimensions: Intelligence and Index (I&I), Calibration and Certification (C&C), and Validation and Verification (V&V), for *trustworthy AI*. *Scenarios engineering* surpasses the features engineering heavily dependent on machine learning and data science to the "scenarios" of problems, requirements, behaviors, data, and solutions for omniscient, calibrated, certified, valid, and verifiable AI systems and solution development.

AI Focus presents the article "Data-Driven Predictive Maintenance" by João Gama et al. They discuss the demand for *predictive maintenance* for Industry 4.0, 5G networks, and IoT networks. The roles of predictive maintenance for predicting fault, remaining useful life, and the root cause of anomalies are discussed. They conclude the article by discussing the future directions, such as digital virtualization (e.g., digital twins), for predictive maintenance.

I am pleased to announce the new column, AI Insight, edited by Professor Usama Fayyad. In the first article for this column, Usama shares his expert opinions on "The Secrets of Data Science Deployments" for a successful production. He calls for attention to various issues of deployment in production, including establishing trust with business users and stakeholders, respecting constraints in model deployment, and appreciating the costs and return on investment in deploying data science models in production. He encourages us to think beyond just data problems, take small steps to build continuous and measurable progress, collect the right data, build the right infrastructure, and develop the right data, science models. His abovementioned thoughts recall my early advocate of *domain-driven data mining* and *actionable knowledge discovery and delivery*. Classic data-driven methodologies and practices are usually ineffective in generating actionable solutions. They have to be transformed toward being real-life problems-driven, involving domain knowledge and factors in modeling, developing

1541-1672 © 2022 IEEE
Digital Object Identifier 10.1109/MIS.2022.3197949
Date of current version 12 September 2022.

models capable of operations from input processing to action recommendation, evaluating results in terms of subjective and objective measures, and delivering high-utility actions for business users.

In addition, the department AI and Cyber-Physical-Social Systems presents the article titled "Metaverses and DeMetaverses: From Digital Twins in CPS to Parallel Intelligence in CPSS." Finally, the department Affective Computing and Sentiment Analysis presents the article titled "Multiscale 3D-Shift Graph Convolution Network for Emotion Recognition from Human Actions," which shares new thoughts on the initial discussion on cyber-physical-social systems (CPSS), concerned in this department since its establishment.

NON-IID LEARNING AND FEATURE ARTICLES

In this July/August issue, I select six feature articles that are clustered on the theme "Non-IID Learning." This theme carries forward the discussion on "Non-IID Federated Learning" in the March/April issue, whose focus was on handling the non-IIDness in federated learning.

As discussed in my editor's perspective article "Beyond i.i.d.: Non-IID Thinking, Informatics, and Learning," non-IIDness is an intrinsic characteristic of any real-life system, behavior, and data. In contrast to *IIDness*, i.e., independence and identical distribution, *non-IIDness* refers to the couplings and interactions and heterogeneities within and between variables, objects, and subsystems of a system. They may be further presented in various forms, types, structures, distributions, relations, granularities, modalities, and hierarchies, etc. Existing analytical and learning systems often only capture some of these aspects and characterize them by specific means, such as dependence, correlation, and association. Such well-defined mathematical tools cannot capture other couplings, interactions, and heterogeneities.[a]

Arguably, existing learning systems, including Bayesian learning, transfer learning, federated learning, and deep neural learning, do not capture comprehensive non-IIDnesses in complex real-life problems. The widely assumed principle of decoupled and disentangled representation and learning in deep learning simplifies the usually comprehensive non-IIDnesses in complex data. This results in typical problems in deep learning, including network vulnerability, over- or under-fitting, and poor adaptability to evolving data and settings (e.g., over in-samples and out-of-samples, open domains, and open tasks).

In this issue, six articles address various aspects and scenarios of data that may present certain non-IIDnesses. First, the article "Contribution- and Participation-Based Federated Learning on Non-IID Data" addresses the impact of statistical client heterogeneity by allocating different client distribution ratios and then aggregating them.

In the article "Bayesian Optimization for Expensive Smooth-Varying Functions," the authors go beyond the uniform smoothness assumption commonly taken in Bayesian optimization to consider the objective function by a set of local and global Gaussian process (GP) models for various regions of input spaces. The posteriors of local and global GP models estimate the mean and variance of any input sample.

Furthermore, the article "Fast Approximate Multioutput Gaussian Processes" continues the discussion on GP models but addresses the computational cost of large samples. It reduces the complexity of generating approximate GPs by approximating the covariance kernel with eigenvalues and functions. Their method can regress over multiple outputs with correlations.

In the article "FD-LSTM: A Fuzzy LSTM Model for Chaotic Time-Series Prediction," a deep fuzzy long short-term memory network (LSTM) combines LSTM and type-2 fuzzy logic to learn the high-order uncertainty in time series forecasting.

In addition, the article "DualNER: A Trigger-Based Dual Learning Framework for Low-Resource Named Entity Recognition" addresses named entity recognition with a small amount of annotated data by selecting effective other entities as "triggers." The identified effective triggers are then used to learn a dual learning framework for low-resourced named entity recognition.

Finally, the article "Intelligent Pandemic Surveillance via Privacy-Preserving Crowdsensing" introduces a differential privacy model to obtain insights from crowdsensing regions of different sizes while protecting the data privacy.

CONCLUDING REMARKS

It is understandable that the methods introduced in the feature articles may not sufficiently address non-IID challenges or were not designed for non-IID learning. I hope this theme encourages your thinking and research on fundamental theories and tools for handling various non-IIDnesses in real-life problems, systems, behaviors, and data.

I hope you enjoy this issue.

[a]Interested readers may refer to https://datasciences.org/non-IID-learning for more information about non-IIDness and non-IID learning.