データ駆動型物質・材料研究の諸問題：限られたデータの壁を乗り越える

（情報・システム研究機構 統計数理研究所 [1]・物質・材料研究機構 [2]）○吉田 亮 [1,2]

Materials Informatics: Overcoming Barriers of Limited Data

([1]*The Institute of Statistical Mathematics, Research Organization of Information and Systems*, [2]*National Institute for Materials Science*) ○Ryo Yoshida[1,2]

Keywords : Materials Informatics; Inverse Design; Small Data; Transfer Learning; Functional Data Analysis

Material informatics (MI), a new form of materials research that combines materials data with data science, is gaining traction. MI applies machine learning (ML) to predict new materials with innovative properties and their fabrication methods from a vast design space. Over the past few years, MI technologies have spread rapidly in various areas of materials research, and many new materials have been discovered [1,2]. However, the application of ML in materials science is lagging behind that in other research areas. Needless to say, data is the most important resource in data-driven science. However, efforts toward creating a comprehensive database of material properties to enable data-driven research have been insufficient. In this talk, I will describe some key technologies of ML to overcome the big hurdle of limited data.

ML techniques called transfer learning or domain adaptation have the great potential to break the barrier of limited data [2,3,4,5]. For a given task to be predicted from a limited supply of training data, a set of models on related tasks are pre-trained using an enough amount of data, which capture common features relevant to the target task. Re-purposing such features on the target task brings an outstanding prediction performance even with exceedingly small data as if highly experienced human experts can perform rational inferences even on considerably less experienced tasks.

The second topic focuses on ML techniques from adaptive experimental design. Any ML models are interpolative in nature, and their prediction capability is no longer applicable in regions where no data are available. However, the ultimate goal of materials science is the discovery of truly innovative materials, which would reside in yet-unexplored material space where no one has gone before. A promising solution to this problem is the integration of computer/physical experiments into a ML system through experimental design techniques such as Bayesian optimization [5].

I also show the potential of supervised learning for ultra-high-dimensional or functional-type output variables. In machine learning of material data, the output variable is often given as a function (Figure 1). For example, when predicting the optical absorption spectrum of a molecule, the output variable is given as a spectral function defined in the wavelength domain. Alternatively, in predicting the microstructure of a composite material, the output variable is given as an image from an electron microscope, which can be represented as a two- or three-dimensional function in the image coordinate system. Here we consider a unified framework to handle such multidimensional or functional output regressions, which are applicable to a wide range of predictive analyses [6]. Of particular interest here is the mechanism of the high tolerance of the functional output regression to limited data. As shown, the present method predicting the whole function directly has statistical mechanisms closely related multitask learning; multiple related tasks are learned simultaneously, allowing the model to recognize common mechanisms among target tasks and consequently improve the prediction accuracy of each task. It is demonstrated that a similar learning mechanism is expected to work in regression with high-dimensional output variables.

データ駆動型研究における最も重要な資源はいうまでもなくデータである．しかしながら，物質・材料科学ではデータ駆動型研究に資する体系的なオープンデータを創出しようという動きは極めて低調である．また，革新的な材料の周辺にはデータは存在しない．当然ながら，データが存在しない未踏領域ではデータ科学の“内挿的な予測”の有効性は失われる．したがって，原理的には戦略なき単純なデータ駆動型アプローチでは，真に革新的な材料の発見には到らない．すなわち，データ駆動型材料研究が抱える問題の本質は，データがないということである．本講演では，限られたデータの壁を乗り越えるための統計的機械学習の方法論として，転移学習，適応的実験計画法による実験・シミュレーション・機械学習の融合，温度依存物性，分光スペクトル，材料微細組織等の関数出力変数の予測等に関する解析手法や関連分野の取り組みを紹介しながら，克服すべき技術的課題や将来展望を論じる．

[1] Liu et al. Machine learning to predict quasicrystals from chemical compositions. Advanced Materials 33, 2102507 (2021).
[2] Wu et al. Machine-learning-assisted discovery of polymers with high thermal conductivity using a molecular design algorithm. npj Computational Materials 5, 66 (2019).
[3] Yamada et al. Predicting materials properties with little data using shotgun transfer learning. ACS Central Science 5, 1717–1730 (2019).
[4] Ju et al. Exploring diamondlike lattice thermal conductivity crystals via feature-based transfer learning. Physical Review Materials 5, 053801 (2021).
[5] Hayashi et al. RadonPy: automated physical property calculation using all-atom classical molecular dynamics simulations for polymer informatics. npj Computational Materials 8, 222 (2022).
[6] Iwayama et al. Functional output regression for machine learning in materials science. Journal of Chemical Information and Modeling 62, 4837–4851 (2022).