Opportunities and Challenges for Inorganic-Material Informatics from a View Point of Big Data Analytics

Yuzuru Tanaka

Data Science G, MI^{2}I

Recent development of big data core technologies including analysis algorithms and high performance data management and analysis platform technologies, together with the development of automatic measurement instruments and/or large-scale high-performance computer simulation technologies, are currently strongly promoting the paradigm shift from mission-driven research to data-driven research in varieties of domain sciences, which is gradually allowing us to conduct scientific research studies completely in cyber worlds after having obtained all the required data sets, or through the real-time receiving of data streams. This trend which is sometimes called in-silico science has been motivating "X" informatics for varieties of "X" sciences such as bio, biomedical, chemical, geo, brain, cosmological, meteorological, pharmaceutical, and material sciences, where the typical approach is to apply brute force analysis power to both experimental and simulation big data for the large-scale exploration to discover new objects or rules, and/or for the precise prediction of the complex dynamic phenomenea. In-silico sciences will further allow us to easily share and exchange not only data sets but also analysis and visualization tools and services, analysis scenarios, and meta knowledge about them, and will definitely lead us to what we call open science, or further to what we call citizen science.

Bioinformatics has made the first big success among data-driven sciences to encourage other sciences to follow. Personalized medicine and material informatics are example followers. However, their researchers are gradually recognizing the difficulties to fill in the gap between varieties of available data analysis methods and the goals to find out new meaningful personalized treatments or new functional materials.

This gap has three major causes. The first one is the mismatch between the issues and the methods in interdisciplinary research collaborations. Big data analysis research in CS focuses on the cutting-edge new algorithms; currently they are, for example, deep learning and Bayesian optimization. For computer scientists to publish their research collaboration results in CS community, they are often biased to pick up such cutting-edge methods, instead of applying legacy methods such as SVR and item-set and association-rule mining. Each method has its own prerequisites which may often mismatch with the characteristics of our current issues to solve.

In these data-driven sciences, most of the target systems are complex systems of systems in which more than one subsystem with different mechanisms interact with each other, and each of them is also a heterogeneous system, i.e., a mixture of more than one subsystem following either different mathematical models or the same model with different parameter values. In the machine learning of such a system, the learning data set inherently consists of more than one subset that follow different mathematical models or the same model with different parameter values. It is necessary to appropriately segment the learning data set into homogeneous subsets before applying the machine learning separately to each subset. Such segmentation is generally not an easy task. Furthermore, the size of each homogeneous data subset may often become too small for statistically meaningful analysis. Exploratory visual analytics may be one of the solutions to this issue [1].

Each existing large-scale database of inorganic natural materials is also a mixture of different types of materials consisting of different atoms arranged in different structures. The total number of the learning

data for a certain type of inorganic natural materials for which we can assume the same physical model for simulation and/or the same regression model for analysis may be in the order of 10^3, or 10^4 at most, which is definitely too small for machine learning, and definitely not sufficient for the deep learning that typically requires more than 10^6 data.

Besides the first cause of the gap, i.e., the heterogeneity of the learning data set and the comparatively small size of each homogeneous data subset, it is often difficult to define a sufficiently large number of appropriate explanatory variables, i.e., descriptors, in providing the learning data set through measurement and/or simulation. In bioinformatics, "genome" constitutes substantial portion of explanatory variables. In material informatics, we also need its counterpart, i.e., "materials genome". For proteins and peptides, a web server called PROFEAT computes structural and physicochemical features from amino acid sequence to systematically define a sufficient number of explanatory variables. It is a challenge, especially in inorganic material informatics, to systematically define a sufficiently large number of appropriate explanatory variables, i.e., inorganic materials genome [2, 3].

In order to increase the size of each homogeneous subset of the learning data set, we may focus more attention on specific families of artificially designed inorganic materials than on natural ones, where a family means a category of materials sharing the same type of structures. Examples may include those with amorphous structures, those with higher-order crystal structures, and organometallic materials with varieties of modifications. Such higher-order nanostructures and/or mesoscopic structures may increase not only the design parameters but also the value space spanned by these design parameter variables. These design parameters may work as explanatory variables of the learning data set, which also has some objective variables whose values are either calculated by the simulation based on the first-principle-calculation modeling of the artificial materials, or provided by material databases showing the relations among related physical properties of the involving atoms, crystal structures, and the functional properties. We can compute only a sufficiently large finite number of simulations to calculate some functional properties of our concern. These functional properties of the materials may include conductivity, magnetic property, optical property, interfacial activity, catalytic activity, and bulk modulus. The machine learning for the regression using the simulation result as the learning data set will estimate the values of such physicochemical properties either for arbitrary value combinations of explanatory variables for which the simulation is still missing, or for their possible value combinations obtained by the simulation only based on energy stability without calculating functional properties [2, 3].

It is not always possible to mathematically model the total system with all the physicochemical and structural parameters taken into account as explanatory variables for estimating some functional properties of our concern. The original idea of machine learning was to give a solution to this problem. Instead of assuming the knowledge about the underlying mechanism of the total system, it uses the observation records of the relation between a sufficiently large set of aspects and each functional property of the system as its learning data set to estimate this functional property value for an arbitrary new value combination of aspects. The success of machine learning heavily depends on the quality and the quantity of such aspects of the target system. Each aspect defines explanatory variables as parameters of its mathematical modeling [4]. In the simplest case, an aspect defines a single explanatory variable.

Aspect modeling is different from the total-system modeling. It may use a simple model that may explain the specified aspect of the system [4]. In naive application of machine learning to materials data, some material properties become difficult to estimate accurately. Material properties such as lattice constant and magnetic moment can be accurately estimated from simple descriptors, i.e., explanatory variable, using basic machine learning methods. However, in the experiments, machine learning did not work well to

estimate the material bulk modulus (the resistance to compression of the material). After adding new explanatory variables such as bond type, energy difference in compression and expansion, and density for the aspect modeling of the material bulk modulus, and calculating, for each record in the learning data set, the values of these added explanatory variables through the simulation of this aspect modelling, the bulk modulus could be well estimated [4, 5]. The energy difference in compression and expansion is the dynamic behavior aspect of the material, which was not considered in the original simulation data focusing on energy-stable structures. The density can be calculated after analyzing the geometrical structure of each simulation result, and is not used as a parameter variable of the simulation. Here we introduce new terminology, genotypic and phenotypic explanatory variables. Genotypic variables are simulation parameter variables, while phenotypic variables are explanatory variables characterizing the phenotypic structures or behaviors of the simulated results. Examples of phenotypic variables include density, fractal dimension, and persistent homology indices. We need to find out much more different types of phenotypic variables.

Some aspect of our concern may be defined as a function of already defined explanatory variables. Depending on the types of machine learning, such an aspect may require the explicit introduction of a new explanatory variable as a derived variable, i.e., a function of other variables. In support vector machine regression, derived variables defined as polynomials of other explanatory variables need not be explicitly introduced as new explanatory variables. They are implicitly considered by the algorithm if necessary. However, such a derived variable as x/y or log x should be explicitly introduced as a new explanatory variable. Some indices obtained as analysis results such as cluster ids or pattern ids may sometime work as new explanatory variables for further segmentation and analysis. We call such explanatory variables marker variables or, simply, markers [1].

It should be noticed that the design of appropriate explanatory variables and the process of segmentation and analysis are both by their nature exploratory processes. This implies the importance of the development of an integrated exploratory visual analytics platform for data-driven sciences [1]. A further shift toward open science requires not only the sharing of platform systems, but also a shared repository of data sets, analysis and visualization tools and services, analysis scenarios, and meta knowledge about them in reusable forms. Meme media and meme pool architectures as well as their web-based implementation Webble World will be able to answer these requirements, and provides a potential middleware framework for end-to-end in-silico exploratory material science and engineering.

1) Y. Tanaka, J. Sjöbergh, K. Takahashi:: A Need for Exploratory Visual Analytics in Big Data Research and for Open Science. IV 2016: 261-270
2) K.Takahashi, Y.Tanaka: Material Informatics: a journey towards material design and synthesis, Dalton Trans. 2016, 45, 10497-10499
3) K. Takahashi, Y. Tanaka: Material synthesis and design from first principle calculations and machine learning, Computational Materials Science, Vol.112, Part A, pp.364-367, Feb. 2016.
4) K. Takahashi, Y. Tanaka: Role of descriptors in predicting the dissolution energy of embedded oxides and the bulk modulus ofoxide-embedded iron, Physical Review B 95 (1), 014101 (accepted)
5) K. Takahashi and Y. Tanaka: Unveiling the material genome of amorphous carbon, Physical Review B (accepted)