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

Yuzuru Tanaka
tanaka@meme.hokudai.ac.jp
(Graduate School of Information Science and Technology, Hokkaido University)

1. Opportunities for Data-driven Sciences

While “big data” in general is characterized by 3V, i.e., the volume, the velocity and the variety of the target data set and/or data stream, by 4V, adding the veracity of data, or by 5V, adding the value of the analysis result, “big data” in applications, especially in cutting-edge science, symbolizes the paradigm shift from mission-driven research to data-driven research, where the volume may not be the major property of the target data set in the current situation. 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 this paradigm shift 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 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.

2. Challenges for Data-driven Sciences

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 two major causes.

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. Personalized medicine aims to find out a personalized treatment that works best for a specific patient, but not necessarily well for the others. The learning data set of patients is inherently a mixture of different types of patients with different chemo-responses. 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 small for machine learning, and definitely not sufficient for the deep learning.

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 sufficient number of appropriate explanatory variables 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 sufficient number of appropriate explanatory variables, i.e., inorganic materials genome.

3. Proposed Action Plan for Inorganic Material Informatics from a Computer Scientist’s View Point

In order to increase the size of each homogeneous subset of the learning data set, we may focus more attention on artificial inorganic materials than on natural ones. Examples may include those with amorphous structures and those with higher-order crystal structures of atom clusters. 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. An amorphous material, for example, may introduce two more design parameters, i.e., the average and the variance of its crystalline diameters. A super crystal of atom clusters may introduce the design parameters of both each atom cluster and the super crystal structure. These design parameters may work as explanatory variables of the learning data set, which may be provided by the simulation based on the first-principle-calculation modeling of the artificial materials and by databases of related physical properties of the involving atoms and crystal structures. 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 for arbitrary value combinations of explanatory variables for which the simulation is still missing.

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. 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. 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 [1]. 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.

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 linear-regression machine learning, derived variables defined as linear combinations 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 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.

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. 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 [2] as well as their web-based implementation Webble World will answer these requirements.

Reference

1) K. Takahashi and Y. Tanaka, “Material synthesis and design from first principle calculations and machine learning,” Computational Materials Science, vol. 112, pp. 364–367, 2016.

2) Y. Tanaka, Meme Media and Meme Market Architecture. Piscataway; NJ; USA: IEEE Press, 2003.