Journal of of Materials Engineering and Applications
Extended Abstract
2018
Vol. 2, Iss. 1

12th Nanotechnology Products and Summit, November 24-25, 2016 Melbourne, Australia- A Statistical Learning Framework for Materials Science: Application to Elastic Moduli of k-nary Inorganic Polycrystalline Compounds

Adegoke Sarajen
Simon Fraser University Saudi Arabia

Materials scientists increasingly employ machine or statistical learning (SL) techniques to accelerate materials discovery and design. Such pursuits benefit from pooling training data across, and thus being able to generalize predictions over, k-nary compounds of diverse chemistries and structures. This work presents a SL framework that addresses challenges in materials science applications, where datasets are diverse but of modest size, and extreme values are often of interest. Our advances include the application of power or Hölder means to construct descriptors that generalize over chemistry and crystal structure, and the incorporation of multivariate local regression within a gradient boosting framework. The approach is demonstrated by developing SL models to predict bulk and shear moduli (K and G, respectively) for polycrystalline inorganic compounds, using 1,940 compounds from a growing database of calculated elastic moduli for metals, semiconductors and insulators. The usefulness of the models is illustrated by screening for superhard materials. In recent years, first-principles methods for calculating properties of inorganic compounds have advanced to the point that it is now possible, for a wide range of chemistries, to predict many properties of a material before it is synthesized in the lab1. This achievement has spurred the use of high-throughput computing techniques as an engine for the rapid development of extensive databases of calculated material properties. Such databases create new opportunities for computationally-assisted materials discovery and design, providing for a diverse range of engineering applications with custom tailored solutions. But even with current and near-term computing resources, high-throughput techniques can only analyze a fraction of all possible compositions and crystal structures. Thus, statistical learning (SL), or machine learning, offers an express lane to further accelerate materials discovery and inverse design. As statistical learning techniques advance, increasingly general models will allow us to quickly screen materials over broader design spaces and intelligently prioritize the high-throughput analysis of the most promising material candidates.

One encounters several challenges when applying SL to materials science problems. Although many elemental properties are available, we typically do not know how to construct optimal descriptors for each property, over a variable number of constituent elements. For instance, if one believes that some average of atomic radii is an important descriptor, there are many different averages, let alone possible weighting schemes, that one might investigate. This challenge may be reduced by placing restrictions on the number of constituent elements or types of chemistries or structures considered, but such restrictions reduce the generalizability of the learned predictor. Materials science datasets are often also smaller than those available in domains where SL has an established history. This requires that SL be applied with significant care in order to prevent over-fitting the model. Over-fitting leads to predictions that are less generalizable to new data than anticipated, such that predictions are less accurate

Note: This work is partly presented at 12th Nanotechnology Products and Summit, November 24-25, 2016 Melbourne, Australia.

than expected. At the same time, smaller datasets challenge us to use the available data as wisely as possible. This may include leveraging observations related to the smoothness of the underlying physical phenomenon, and the use of an appropriate risk criterion, rather than partitioning the available data into distinct training and test sets. For SL to have the greatest impact on materials discovery and design, we must pursue techniques that make maximal use of the available data. This requires approaches that are capable of systematically pooling training data across, and are thus capable of generalizing predictions over, k-nary compounds of diverse chemistries and structures.

The successful application of SL requires the selection of an appropriate set of descriptor candidates. In materials science problems, the candidates must be capable of both "uniquely characterizing" a diverse range of compounds, and sufficiently explaining the diversity of the phenomenon being learned. Thus, the selection of descriptor candidates is a crucial and active field of investigation within materials science, as the field endeavors to develop general models with high predictive accuracy. Previous work in materials science has included both categorical descriptors and continuous descriptors. Although both types of descriptors may be legitimately used in SL, special care should be taken when using categorical descriptors, as each such descriptor essentially (i.e., unless there is sufficient smoothing across cells) partitions the space of compounds into disjoint cells, which quickly increases the degrees of freedom and thus the risk of over-fitting the model.

SL applications should always include descriptor candidates suggested by known, scientifically relevant relationships. But in order to construct models that accurately generalize across diverse datasets, such candidates will typically need to be augmented with additional descriptor candidates, capable of bridging across the simplifying assumptions that divide less generalizable models. Without these additional candidates, attempts to learn more general models will be stifled, as it will be impossible to discover new, unexpected relationships. Here we introduce the use of Hölder means, also known as generalized or power means, as an ordered approach to explicitly constructing descriptor candidates from variable length numeric lists. Hölder means describe a family of means that range from the minimum to maximum functions, and include the harmonic, geometric, arithmetic, and quadratic means. This paper advances previous work by constructing descriptor candidates as Hölder means, which, to the best of our knowledge, has not previously been done in the field of materials science.

Having discussed the construction of descriptor candidates, we now introduce gradient boosting machine local polynomial regression, which is a SL technique that we developed to leverage the available data as wisely as possible. Energy minimization problems often enforce smoothness in the functions mapping useful descriptors to outcomes. Statistical learning techniques may exploit such smoothness, when present, in order to produce models that are as accurate as possible for a fixed amount of training data; such considerations are more important when working with smaller training datasets than with larger datasets. GBM-Locfit utilizes multivariate local polynomial regression, as implemented in Locfit, within a gradient boosting machine framework. Local polynomial regression performs a series of weighted regressions within a moving window, with a weight function that gives greatest weight to observations near the center of the window, producing a smooth curve that runs it.

Note: This work is partly presented at 12th Nanotechnology Products and Summit, November 24-25, 2016 Melbourne, Australia.