Computational Materials Science 98 (2015) 170–177

Contents lists available at ScienceDirect

Computational Materials Science

journal homepage: www.elsevier.com/locate/commatsci

Evaluation of machine learning interpolation techniques for prediction of physical properties

Eve Bélisle [a], Zi Huang [a], Sébastien Le Digabel [c], Aïmen E. Gheribi [b,*]

[a] School of Information Technology and Electrical Engineering, The University of Queensland, Brisbane QLD 4072, Australia
[b] CRCT Center for Research in Computational Thermochemistry, Department of Chemical Eng., École Polytechnique de Montreal (Campus of Université de Montréal), Box 6079, Station Downtown, Montréal, Québec H3C 3A7, Canada
[c] GERAD and Department of Mathematics and Industrial Eng., École Polytechnique de Montréal, Succ. Centre-ville, Montréal, Québec H3C 3A7, Canada

ARTICLE INFO

Article history:
Received 25 June 2014
Received in revised form 14 October 2014
Accepted 17 October 2014

Keywords:
Superalloys
Database
Gaussian process
Neural network
Quadratic regression
Physical properties
Computational dependence

ABSTRACT

A knowledge of the physical properties of materials as a function of temperature, composition, applied external stresses, etc. is an important consideration in materials and process design. For new systems, such properties may be unknown and hard to measure or estimate from numerical simulations such as molecular dynamics. Engineers rely on machine learning to employ existing data in order to predict properties for new systems. Several techniques are currently used for such purposes. These include neural network, polynomial interpolation and Gaussian processes as well as the more recent dynamic trees and scalable Gaussian processes. In this paper we compare these approaches for three sets of materials sciences data: molar volume, electrical conductivity and Martensite start temperature. We make recommendations depending on the nature of the data. We demonstrate that a thorough knowledge of the problem beforehand is critical in selecting the most successful machine learning technique. Our findings show that the Gaussian process regression technique gives very good predictions for all three sets of tested data. Typically, Gaussian process is very slow with a computational complexity of typically n^3 where n is the number of data points. In this paper, we found that the scalable Gaussian process approach was able to maintain the high accuracy of the predictions while improving speed considerably, make on-line learning possible.

1. Introduction

Through the years, various machine learning techniques have been employed to fit sets of known data associated with certain properties in order to predict these properties on sets of unknown data. The "No Free Lunch theorem" was introduced in 1997 by Wolpert and Macready [1], stating that for every optimization problem, there is no perfect algorithm. For a given problem for which an approach works well, there exists another problem for which the same method fails miserably. This paper aims at comparing different machine learning techniques for predicting properties of different types of data. Our focus is on material science data of molten oxides systems collected from the literature. It is important for material science engineers to know the physical properties of such systems in order to design new materials, or improve the current processes.

Currently, there are a variety of machine learning techniques for predicting a function $f(x)$ given x. Polynomial interpolation was one of the first to be developed [2], and is still a very popular method in fields such as digital photography and image re-sampling as well as for scientific data. Gaussian processes (GPs) were introduced in the 1940s [3], but it is only in 1978 that they were employed to define prior distributions over functions [4]. More recently, with the introduction and increasing popularity of neural networks with back propagation, Gaussian processes started to be used for supervised machine learning [5] and for regression problems [6]. In the last few years, various attempts have been made to improve known approaches, in particular by the group of Robert B. Gramacy at the University of Chicago, with the introduction of treed Gaussian processes [7] and dynamic trees [8]. In 1996, Radford Neal showed that a Bayesian neural network with a Gaussian prior on individual weights with an infinite number of hidden nodes converges to a GP [9].

In this work, we perform a comparative study of the predicting power of six of the most popular and emerging machine learning techniques. The different techniques are tested on datasets from the materials science industry: molar volume (MV), electrical

* Corresponding author.
E-mail addresses: e.belisle@uq.edu.au (E. Bélisle), uqzhuang@uq.edu.com (Z. Huang), sebastien.le-digabel@polymtl.ca (S. Le Digabel), aimen.gheribi@polymtl.ca (A.E. Gheribi).

http://dx.doi.org/10.1016/j.commatsci.2014.10.032
0927-0256/

Nomenclature

δ	Kronecker delta	n	number of training (experimental) points
μ	mean	N_G	Gaussian Noise
κ	electrical conductivity	NS	Nash–Sutcliffe model efficiency
σ	variance	RMSE	root mean square error
D	number of dimensions	T	temperature
GP	Gaussian process	w	width of a Gaussian kernel
Ms	Martensite start temperature		
MV	molar volume		

conductivity (EC) and Martensite temperature (Ms), respectively consisting of smooth, nonsmooth and noisy (several local minima in a small domain) data. We consider data on molar volume to be smooth because the theory tells us that it should vary almost linearly and also because the experimental datasets are in good agreement at equal composition and temperature. For the electrical conductivity, the data is very scattered and that is why we consider it to be nonsmooth. As for Ms, we consider the data noisy because here we are omitting to include certain influential parameters such as the fine austenite grain size [10] and are considering uniquely the initial composition. We wish to demonstrate how a thorough knowledge of the system as well as machine–human interactions can improve the quality of the predictions. Stry et al. compared the quadratic and linear interpolation applied to the numerical simulation of crystal growth [11]. They found that a custom quadratic approach developed by them gave more accurate results with smaller computational time. Ghosh and Rudy found an improvement of the relative error of reconstructed versus measured epicardial potentials of Electrocardiographic Imaging when using a quadratic interpolation instead of linear one [12]. Skinner and Broughton published their work on neural networks applied to material science, and compared different methods for finding the weights of feed-forward neural networks [13]. In the present paper we have added comparisons with more recent techniques: linear and quadratic interpolation, neural network, Gaussian processes (GP), and dynamic trees. We also include a comparison with a new strategy, the scalable Gaussian process regression (SGP) [14] that was developed to speed up Gaussian process regression while maintaining an acceptable prediction error. This was motivated by the idea to introduce physical properties as one of the possible parameters inside the FactOptimal module of the FactSage software. FactSage is a software system that was created for treating thermodynamic properties and calculations in chemical metallurgy [15]. It is used today all over the world by more than 400 universities and companies in the domain of material chemistry. It contains various modules allowing users to perform a wide variety of thermochemical calculations [16]. The FactOptimal module [17–19] allows one to find the best set of conditions given constraints while optimizing chosen properties. The program uses the NOMAD derivative-free solver [20] to find the best parameters. For example, given chemical system (ex. $x_1\mathrm{C} + x_2\mathrm{Mn} + x_3\mathrm{Si} + x_4\mathrm{Cr}$), one may wish to find the values of chemical compositions (x_i) that would give an equilibrium temperature of around 275 °C. To do so, NOMAD tries different combinations of compositions (x_i), obtaining the corresponding value of temperature from FactSage until, hopefully, an optimal solution is found. The idea to introduce material properties as possible constraints or as values to be optimized requires the use of a machine learning tool to predict these properties. Because a large number of predictions are performed during a FactOptimal run, the computational time to make these predictions is of great importance. Furthermore, we wish to make on-line learning possible, as it may be the case that new experimental data is fed dynamically into the learning database.

Making predictions on the Martensite start temperature is not a new domain. Some authors use a neural network model with good results [21,22]. A thermodynamic framework [23] or a purely empirical approach [24,25] have also been studied. Soumail et al. in 2006 [26] compared these methods and concluded that although the thermodynamic approach provides satisfying results, there is a strict limitation in the query points, based on the fundamental assumptions upon which the model was based. They observed that the neural network approach performs just as good as others but with a higher amount of outliers or wild predictions, therefore they recommended the use of a Bayesian framework. Using a Bayesian GP model, very accurate predictions were obtained for the prediction of austenite formation (Martensite is formed in carbon steels when cooling austenite) [27].

An empirical model [28] and a combined model with quantum chemical molecular dynamics and kinetic Monte Carlo method [29] were applied to predict electrical conductivity. Both models are developed specifically for electrical conductivity and would require extensive work to be adapted to predict other physical properties. To the best of our knowledge, all published material on Ms and electrical conductivity prediction discuss their results in terms of prediction accuracy and no report is given on the computational time.

In the following sections we first provide a description of the databases that were employed for this research, then briefly describe each interpolation technique. Then we present results in terms of computational time and accuracy. We then discuss the results and make recommendations on the use of each method depending on the type of problem.

2. Materials data

For this work, we have access to three databases of experimental points collected from the literature. The database employed for molar volume predictions has 2700 data points ($n = 2700$), with various compositions in mole percent on 10 dimensions ($D = 10$), temperature in Kelvin and an associated molar volume value in cubic centimeters per mole. The electrical conductivity database consists of approximately 9300 data points with compositions in mole percent over 10 dimensions, temperature (T) in Kelvin ($D = 10$) and an associated electrical conductivity (EC) value in Siemens per meter. For both the MV and EC databases, the materials are insulating oxides, therefore EC refers to the ionic conductivity. The Martensite start temperature (Ms) database consists of approximately 1100 data points with composition values in weight percent on 15 dimensions ($D = 15$) and an associated Ms value in Kelvin. The main element, Fe, is not used in the regressions. Table 1 gives the range of compositions of each database.

Some physical properties can be measured with reasonable accuracy, therefore there is very little discrepancy between the different data sources. Moreover, certain properties have a quasi linear dependence with the constituents chemical compositions, while others may have a more complex dependence on compositions and