Journal of System Simulation

Volume 30 | Issue 10 Article 30

1-4-2019

Error Estimation for Material Simulation Data Based on Hybrid Learning Algorithm

Wang Juan
1. Computer Network Information Center, Chinese Academy of Sciences, Beijing 100190, China; ;2. University of Chinese Academy of Sciences, Beijing 100049, China;

Xiaoyu Yang
1. Computer Network Information Center, Chinese Academy of Sciences, Beijing 100190, China; ;

Zongguo Wang
1. Computer Network Information Center, Chinese Academy of Sciences, Beijing 100190, China; ;

Ren Jie
1. Computer Network Information Center, Chinese Academy of Sciences, Beijing 100190, China; ;2. University of Chinese Academy of Sciences, Beijing 100049, China;

See next page for additional authors

Error Estimation for Material Simulation Data Based on Hybrid Learning Algorithm

Abstract

Abstract: *In order to obtain high quality material simulation data from Density Functional Theory material calculation software package, a modeling method based on BP neural network was proposed to build model estimating the error of material simulation data. A novel hybrid algorithm combining simple particle swarm optimization algorithm that excludes speed item with BP algorithm, also referred to tsPSO-BP, was proposed to optimize the connection weights of the BP neural network. The hybrid learning algorithm not only makes use of strong global searching ability of the PSO, but also strong local searching ability of the BP algorithm.* The BP neural network model was trained using tsPSO-BP on the dataset of experimental and calculation data of elastic constants for binary alloys with cubic crystal system, and the results show that the prediction accuracy of the error of C_{11}, C_{12} and C_{44} were 88.19%, 87.83% and 88.26%, respectively.

Keywords

error estimation for material simulation data, neural network, particle swarm optimization, BP algorithm, hybrid learning algorithm

Authors

Wang Juan, Xiaoyu Yang, Zongguo Wang, Ren Jie, and Xushan Zhao

Recommended Citation

Wang Juan, Yang Xiaoyu, Wang Zongguo, Ren Jie, Zhao Xushan. Error Estimation for Material Simulation Data Based on Hybrid Learning Algorithm[J]. Journal of System Simulation, 2018, 30(10): 3843-3853.

第 30 卷第 10 期 系统仿真学报© Vol. 30 No. 10
2018 年 10 月 Journal of System Simulation Oct., 2018

基于混合学习算法的材料计算数据误差估计

王娟[1,2]，杨小渝[1]，王宗国[1]，任杰[1,2]，赵旭山[1]
(1. 中国科学院计算机网络信息中心 北京 100190; 2. 中国科学院大学 北京 100049)

摘要: *鉴于低误差的基于密度泛函理论的材料计算模拟数据在新材料设计与发现中的重要性，提出一种基于 BP 神经网络和粒子群优化(particle swarm optimization, PSO)混合学习算法的材料计算数据误差估计建模方法。结合 PSO 的全局搜索和 BP 算法的局部搜索优点，将不含速度项的简化 PSO 算法和 BP 算法相结合，提出一种 PSO 和 BP 混合的学习方法(tsPSO-BP)，用于训练材料计算模拟数据误差估计神经网络模型，并以立方晶系二元合金弹性常数计算模拟数据误差估计为应用实例。*应用结果表明 tsPSO-BP 训练后的弹性常数计算模拟误差预测神经网络模型预测的 C_{11}，C_{12} 和 C_{44} 的计算模拟数据误差的准确率分别达到 88.19%,87.83%和 88.26%。

关键词: 材料计算模拟数据误差估计；神经网络；粒子群优化；BP 算法；混合学习方法

中图分类号：TP183;TP391.9 文献标识码：A 文章编号：1004-731X (2018) 10-3843-11

DOI: 10.16182/j.issn1004731x.joss.201810030

Error Estimation for Material Simulation Data Based on Hybrid Learning Algorithm

Wang Juan[1,2], *Yang Xiaoyu*[1], *Wang Zongguo*[1], *Ren Jie*[1,2], *Zhao Xushan*[1]

(1. Computer Network Information Center, Chinese Academy of Sciences, Beijing 100190, China;
2. University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract: *In order to obtain high quality material simulation data from Density Functional Theory material calculation software package, a modeling method based on BP neural network was proposed to build model estimating the error of material simulation data. A novel hybrid algorithm combining simple particle swarm optimization algorithm that excludes speed item with BP algorithm, also referred to tsPSO-BP, was proposed to optimize the connection weights of the BP neural network. The hybrid learning algorithm not only makes use of strong global searching ability of the PSO, but also strong local searching ability of the BP algorithm*. The BP neural network model was trained using tsPSO-BP on the dataset of experimental and calculation data of elastic constants for binary alloys with cubic crystal system, and the results show that the prediction accuracy of the error of C_{11}, C_{12} and C_{44} were 88.19%, 87.83% and 88.26%, respectively.

Keywords: error estimation for material simulation data; neural network; particle swarm optimization; BP algorithm; hybrid learning algorithm

引言

“材料基因组计划”(Materials Genome Initiative，MGI)[1]的目的是通过材料计算机模拟计算、材料实验和材料数据分析相结合的方法加快新材料的研发。基于密度泛函理论(Density Functional Theory，DFT)[2]的材料计算模拟方法作为支撑 MGI 实施的主要材料计算方法，是分析材料“组分-结构-性质”关系的基础，可靠的材料计算模拟数据为新材料的设计与发现提供了指导。由于多体相互作用的存

收稿日期：2016-09-08 修回日期：2016-12-19;
基金项目：国家自然科学基金 (61472394, 11547177);
作者简介：王娟(1984-)，女，北京，博士，研究方向为人工智能，材料信息学；杨小渝(1968-)，男，北京，博士，研究员，研究方向为科研信息化，材料信息学，科学数据管理。

在，多粒子体系的薛定谔方程无法被精确求解，因此 DFT 材料计算在实际的执行中用到了多种近似，如玻恩-奥本海默近似、交换关联能近似和势能求解近似[3]等，所以材料计算模拟结果存在误差。基于 DFT 的材料计算模拟软件包通常提供一整套参数及参数的取值范围，计算模拟结果的可靠性与输入参数的取值有关联性，需要设置合理的参数取值才能取得较准确的计算结果，因为 DFT 计算模拟输入参数的不同取值，如布里渊区 K 点的取值大小、控制平面波基组的截断能 ENCUT、FFT 格点的精度参数取值等对 DFT 计算模拟结果的影响不同。由于 DFT 计算模拟输入参数的复杂性，如何估计不同输入参数取值时计算模拟结果的误差，从而不需要大量的模拟计算即可找到适合材料体系的输入参数，得到较可靠的模拟结果对加快新材料的设计具有实际应用意义。目前还未见国内外有关 DFT 计算模拟输入参数取不同值时预测模拟计算结果误差的报道。本文用自主研发的高通量材料计算平台 MatCloud[4-5] 构建的材料计算模拟数据库，及搜集到的材料弹性常数性质实验数据为基础，尝试以立方晶系二元合金材料的弹性常数为例，用神经网络方法构建二元合金弹性常数计算模拟数据误差预测模型，估计不同参数组合下弹性常数 DFT 计算模拟数据的误差。本文的研究具有两个创新应用意义：(1)对 DFT 计算模拟参数的设置具有指导意义，计算者用模型推荐的误差较低的参数进行计算，减少寻找最优参数所需要的计算模拟时间和计算资源；(2)增强数据库中存储的 DFT 计算模拟数据的实用性。

因为单隐层前馈神经网络(Single-hidden Layer Feedforward Neural Network， SLFN)具有学习能力强，能够逼近复杂非线性函数，以及能够解决传统参数方法无法解决的问题，所以 SLFN 广泛应用于很多领域；SLFN 已被成功应用在 DFT 计算模拟数据修正、材料弹性常数性质预测中。陈冠华[6]用 BP 神经网络方法建立 DFT 计算模拟数据和实验数据之间的量的关系，对 DFT 计算数据进行修正，缩小模拟数据和实验数据的差值；Ozerdem[7]以合金中元素的原子浓度为输入，成功构建了 Cu–Sn–Pb–Zn–Ni 合金的力学性质神经网络预测模型。众所周知，SLFN 缺乏快速的学习方法，对于目前应用最广泛的学习方法中的 BP 方法，也存在如下缺点[8]：(1)需要很多次迭代，训练速度慢；(2)参数选择比较敏感，必须选取合适的学习率与网络连接权值初始值，如果学习率取值太小，算法收敛很慢，而学习率太大，则算法不太稳定甚至难于收敛；(3)容易陷入局部极小点；(4)容易出现过度拟合的现象。为了解决 BP 神经网络模型中问题(2)(3)(4)，本文将粒子群优化(particle swarm optimization，PSO)算法引入 BP 神经网络的训练中，结合 PSO 和 BP 的优势，用一种 PSO 和 BP 相混合的 SLFN 训练方法，改进 BP 神经网络训练方法。

虽然 PSO 算法具有运算简单，易于实现等优点，但也有进化后期收敛慢，收敛精度不高及不能保证全局最优解的缺点。针对 PSO 算法和 BP 算法的缺点，并结合二者优势，本文将不含速度项的简易 PSO 算法和 BP 算法相结合，使得两者优势互补，并将改进后的神经网络训练方法应用于立方晶系二元合金弹性常数 DFT 计算模拟数据误差估计问题的研究中。

1 PSO 和 BP 混合的 SLFN 训练算法

1.1 SLFN 网络模型

对于含有 N 个样本的训练集 $D=(\boldsymbol{x}_i,\ \boldsymbol{t}_i)$，其中 $\boldsymbol{x}_i=[x_{i1},x_{i2},...,x_{io}]^T\in R^o$，$\boldsymbol{t}_i=[t_{i1},t_{i2},...,t_{im}]\in R^m$，$O$ 和 M 分别表示训练样本的特征数和目标变量的维数。一个具有 $\bar{N}$ 个隐含层节点的 SLFN 的数学模型表示为：

$$\sum_{i=1}^{\bar{N}} f^2(\boldsymbol{w}_i^2\boldsymbol{x}_{ij}^2+\boldsymbol{b}^2)=$$

$$\sum_{i=1}^{\bar{N}}\sum_{k=1}^{O} f^2(\boldsymbol{w}_i^2 f_i^1(w_{ki}^1 x_{kj}+b_i^1)+\boldsymbol{b}^2)=\boldsymbol{t}_j \tag{1}$$

式中 j=1，2，…，N，$\boldsymbol{w}_i^2=[w_{i1}^2,w_{i2}^2,...,w_{im}^2]^T$ 表示第

DOI: 10.16182/j.issn1004731x.joss.201810030

i 个隐含层神经元与输出层神经元的连接权值；$f^2(x)$ 为输出层神经元的激励函数；$\boldsymbol{x}_{ij}^2$ 是第 i 个隐含层神经元的输出；$\boldsymbol{b}^2$ 为输出层神经元的偏置；$f^1(x)$ 为隐含层神经元的激励函数；$\boldsymbol{w}_i^1=[w_{1i},w_{2i},...,w_{oi}]^T$ 为连接第 i 个隐含层神经元的输入权值；b_i^1 表示隐含层第 i 个神经元的偏置。

引入变量 ε 表示神经网络预测值和目标值的差值，SLFN 的优化目标是寻找一组近似最优的参数 $\boldsymbol{W}=(\boldsymbol{w}^1,\boldsymbol{w}^2,\boldsymbol{b}^1,\boldsymbol{b}^2)$，使得预测值和实际值之间的误差平方和最小。优化问题可以表示为：

$$\begin{aligned}&\arg\min_{\boldsymbol{W}=(\boldsymbol{w}^1,\boldsymbol{w}^2,\boldsymbol{b}^1,\boldsymbol{b}^2)} E(\boldsymbol{W})=\arg\min_{\boldsymbol{W}=(\boldsymbol{w}^1,\boldsymbol{w}^2,\boldsymbol{b}^1,\boldsymbol{b}^2)}\|\varepsilon\|^2,\\&\text{s. t.}\ \sum_{i=1}^{\bar{N}}\sum_{k=1}^{O}f^2(\boldsymbol{w}_i^2f_i^1(w_{ki}^1x_{kj}+b_i^1)+\boldsymbol{b}^2)-\boldsymbol{t}_j=\boldsymbol{\varepsilon}_j,\\&\quad j=1,2,...,N\end{aligned}\tag{2}$$

式中，$E(\boldsymbol{W})$ 表示预测值和实际值的误差平方和，$\boldsymbol{\varepsilon}_j=[\varepsilon_{j1},\varepsilon_{j1},...,\varepsilon_{jm}]$ 为第 j 个样本的误差。

BP 神经网络训练方法[9]通过反向误差传播原理不断更新网络的连接权值使得网络的输出和目标值之间的误差平方和达到最小。BP 神经网络训练方法通常采用梯度下降法迭代调整权值 $\boldsymbol{W}$:

$$\boldsymbol{W}_k=\boldsymbol{W}_{k-1}-\eta\frac{\partial E(\boldsymbol{W})}{\partial\boldsymbol{W}}\tag{3}$$

式中，η 表示学习率。式(3)中的 $\boldsymbol{W}$ 的初始值对网络的训练性能敏感，初始值对结果影响较大。

1.2 粒子群算法

粒子群优化是一种基于群体智能的优化方法，粒子群中的每个粒子具有速度和位置两个描述量，粒子在群体中的适应度由该粒子位置坐标对应的优化目标函数值决定，粒子的优劣通过适应度值衡量。PSO 算法首先为群体中的每一个粒子初始化速度和位置，然后通过迭代寻找最优解，在迭代过程中，粒子通过个体极值和全体极值来更新自己的速度和位置。PSO 算法最早是由 Kennedy 和 Eberhart[10]于 1995 年提出来的，其公式表示如下：

$$v_{id}^{t+1}=v_{id}^t+c_1r_1(p_{id}-x_{id}^t)+c_2r_2(p_{gd}-x_{id}^t)\tag{4}$$

$$x_{id}^{t+1}=x_{id}^t+v_{id}^{t+1}\tag{5}$$

式中，$v_{id}\in[v_{\min},v_{\max}]$ 是第 i 个粒子的飞行速度，$v_{\min}$ 和 $v_{\max}$ 是由用户设定的最小和最大常数飞行速度，p_{id} 为该粒子当前搜索到的最优位置，x_{id} 为该粒子在第 d 维空间的位置；p_{gd} 为整个粒子当前搜寻到的最优位置；c_1 和 c_2 为正常数学习因子，通常取 c_1=c_2=2；r_1 和 r_2 为服从(0，1)均匀分布的随机数。

为提高算法平衡局部最优和全局最优的能力，Shi[11]等人引入了动量惯性系数 w 对式(4)进行修改，w 越大，全局搜索能力较强，反之局部搜索能力较强。修改后的公式如下：

$$v_{id}^{t+1}=wv_{id}^t+c_1r_1(p_{id}-x_{id}^t)+c_2r_2(p_{gd}-x_{id}^t)\tag{6}$$

粒子在目标值求解空间内不断根据全局极值和个体极值向着最优解的方向进行搜索，在进化后期收敛速度变慢。胡旺[12]对式(4)中粒子的速度项进行了分析，指出后期收敛慢的原因有很大概率是因为粒子速度大小并不代表粒子能够有效趋近最优解位置，反而可能造成粒子偏离正确的进化方向，从而去掉了式(4)中的速度项，提出了一种简化的 PSO 算法。

1.3 PSO 和 BP 混合的 SLFN 训练算法

为了解决 BP 训练方法的缺点，很多学者将 PSO 算法应用到 BP 神经网络模型的训练中[13-17]。在现有应用 PSO 算法改进 BP 训练方法的文献中，研究者多是基于粒子的“位置”和“速度”两个关键量改进 PSO 算法，用改进的 PSO 算法训练 BP 网络模型参数。这些改进方法以增加 PSO 的全局搜索能力或者加快粒子收敛的速度为目的，在 PSO 算法的基础上增加一些操作算子，如自适应惯性系数变化、变异、杂交等，使得 PSO 算法变得比较复杂，最终的应用具有局限性。文献[14]改进了式(5)中的动量惯性系数动态调节方式以加快粒子的收敛速度，用改进后的 PSO 算法训练 BP 网络连接权值。文献[15]在 PSO 算法中增加交叉和变异操作，以增加粒子多样性，增强算法跳出局部最优的能力。另一种研究思路是结合 PSO 算法和 BP 算法

的各自优势，不增加额外操作运算，用 PSO 和 BP 相混合的操作步骤训练 SLFN 神经网络。文献[16]结合 PSO 算法的全局搜索和 BP 算法的局部搜索优势，提出了混合的 PSO-BP 前馈网络训练算法，首先利用 PSO 算法训练 SLFN 的权值，若进化若干代（如 10 代）后，PSO 的全局极值未发生改变，则利用 BP 算法更新 PSO 的全局极值，以避免粒子在偏离最优值方向上的搜索，降低粒子搜索的 CPU 时间。文献[17]引入粒子多样性变量，当粒子的多样性小于预设的值时，用 BP 算法训练 SLFN，将 SLFN 的权值和阈值作为所有粒子的全局极值，同时对所有粒子的个体极值增加随机扰动，以增强粒子的多样性。

本文针对基本 PSO 算法和 BP 算法存在的不足，并结合二者的优势，将文献[12]中不含速度项的带极值扰动的简化 PSO 算法(extremum disturbed and simple particle swarm optimization, tsPSO)和 BP 算法相结合，提出一种 tsPSO 与 BP 算法相混合的 SLFN 训练算法，将其简称为 tsPSO-BP。

1.3.1 tsPSO-BP 算法思想

简化 PSO 算法和带极值扰动更新的 tsPSO 形式分别为(7)和(8)所示：

$$x_{id}^{t+1} = wx_{id}^{t} + c_1 r_1 (p_{id} - x_{id}) + c_2 r_2 (p_{gd} - x_{id}) \quad (7)$$

$$x_{id}^{t+1} = wx_{id}^{t} + c_1 r_1 (r_3^{t_b > T_b} p_{id} - x_{id}) + c_2 r_2 (p_{gd}^{t_g > T_g} - x_{id}) \quad (8)$$

式中 t_b 和 t_g 分别表示个体极值和全局极值进化停滞步数；T_b 表示个体极值需要扰动的停滞步数阈值；T_g 表示全局极值需要更新的停滞步数阈值。

粒子个体极值和全局极值进化停滞若干步后，有很大概率陷入了局部极值，此时若没有达到最大迭代次数或 SLFN 训练精度不够，需要增加一些操作使粒子摆脱局部极值。本文采用以下策略：在个体极值进化停滞后，对其进行随机扰动，使粒子飞向新的位置，经新的搜索路径寻找最优解；全局极值进化停滞后，以粒子当前的全局最优值为 SLFN 初始连接权值，使用 BP 算法训练 SLFN，在当前全局极值附近搜寻最优解，若使用 BP 算法训练后的 SLFN 的训练误差小于当前 PSO 训练 SLFN 的误差，则用 BP 训练后的 SLFN 连接权值取代当前全局最优值，这使所有粒子迁移到以 BP 算法搜寻到的最优解为中心的领域进行搜索，因此可以更大概率发现最优解，同时为增强对历史信息的记忆能力，避免粒子过度发散而导致搜索时间过长，最差适应度粒子的个体极值替换为历史全局极值。

1.3.2 基于 tsPSO-BP 的 SLFN 训练步骤

Step 1: 输入训练样本数据集。样本数据包括模型输入数据和目标值。

Step 2: 初始化 tsPSO-BP 优化算法相关参数。如惯性系数、个体和全局极值进化停滞步数阈值、BP 算法训练函数和神经元激活函数；

Step 3: 初始化粒子群，每个粒子的位置被初始化为 (0，1) 之间的随机数；

Step 4: 粒子的初始位置记为当前位置 p_b，将 SLFN 的初始连接权值设置为粒子的当前位置 p_b，计算每个粒子的适应度，适应度最好的粒子的位置为全局极值 p_g；

Step 5: 增加粒子迭代次数，调整式(7)、(8)中的动量惯性系数，计算新一代进化粒子的适应度值。若个体极值和全局极值的进化停滞代数小于预设的阈值，用公式(7)更新粒子的位置，产生一组新的粒子。否则对个体极值增加随机扰动，如果全局极值进化停滞步数大于预设的阈值，将适应度最差粒子的个体极值替换为全局极值，同时用 BP 算法训练 SLFN，将全局极值更新为 SLFN 的权值，用公式(8)更新粒子位置，产生一组新粒子；

Step 6: 计算新产生粒子的适应度值；

Step 7: 如果达到最大迭代次数或者训练精度满足预设精度，则转步骤 8，否则转步骤 5；

Step 8: 将粒子全局极值设为 SLFN 的权值，用 BP 算法训练最终的 SLFN；

Step 9: 输出最终训练好的 SLFN 模型。

DOI: 10.16182/j.issn1004731x.joss.201810030

2 材料计算模拟数据误差估计模型

2.1 模型建立思想

不同晶体结构的材料，各向异性程度不一样，根据 Neumann 原则[18]，晶体的对称性不仅表现在结构上，也表现在物理特性上，因此晶体材料的性质必然和其对称性密切相关。材料计算模拟数据误差为计算值和实验值的差值。通常情况下，单质材料许多性质的实验数据容易获取，并且自然界中单质数量较少，因此可借助高通量材料计算平台，如 MatCloud，对所有单质的物性进行模拟计算，并基于实验数据，计算其误差。基于 DFT 的材料计算中，不同的泛函、材料类别，其误差的系统性往往较为明显[19]，不同元素的细致计算参数所导致的误差，其迁移性也较为明显。因此，对于给定的一组计算输入参数，与利用该组参数计算得到的化合物的计算数据误差关联的因素包括：单质的计算模拟数据误差、化合物与单质的结构差异。这种关系一旦确定，参考特定参数组合下单质的计算模拟数据误差和结构信息，便可估算出该参数集合下化合物的计算模拟数据误差。用单质材料的 DFT 计算模拟数据误差及材料结构属性推导化合物材料 DFT 计算模拟数据误差，这种方法在不同体系上的适用性表现不一样。一般而言，与几何构型相关的性质，如弹性常数，表现较好；与电子相关的性质，如能带和介电常数，该方法适用性较弱，因为电子能级误差可控性、预见性较差。

2.2 模型构建

DFT 计算模拟数据误差预测模型建立的过程是以单质和化合物的结构属性及单质的 DFT 计算模拟数据误差为输入变量，建立输入变量和化合物材料 DFT 计算模拟数据误差的关系模型，量化输入变量和输出变量的函数关系。以立方晶系二元合金材料的弹性常数 DFT 计算数据误差估计为例子，本文所考虑的输入变量有 20 个，如表 1 所示，输出为立方晶系 3 个弹性常数分量 C_{11}，C_{12} 和 C_{44} 的计算模拟值误差。立方晶系二元合金材料弹性常数计算模拟数据误差估计神经网络模型如图 1 所示。

表 1 二元合金弹性常数计算数据误差估计模型输入变量
Tab. 1 Predictor variables of the model for estimation error of calculation elastic constants for binary alloys

输入变量	
晶体结构相似度，sim	$sim_{A\text{-}AxBy}$，$sim_{B\text{-}AxBy}$
原子质量，m	m_A， m_B
原子浓度，c	c_A，c_B
原子总数，n	n_{AxBy}，n_A，n_B
键长，bl	$bl\text{-}AxBy_{A\text{-}B}$，$bl\text{-}AxBy_{A\text{-}A}$，$bl\ \text{-}AxBy_{B\text{-}B}$，$bl\ -A_{A\text{-}A}$，$bl\ -B_{B\text{-}B}$
单质体积计算模拟值，V	V_A，V_B
单质单点能计算模拟值，E	E_A，E_B
单质弹性常数计算误差，sim-error	sim-error_A，sim-error_B

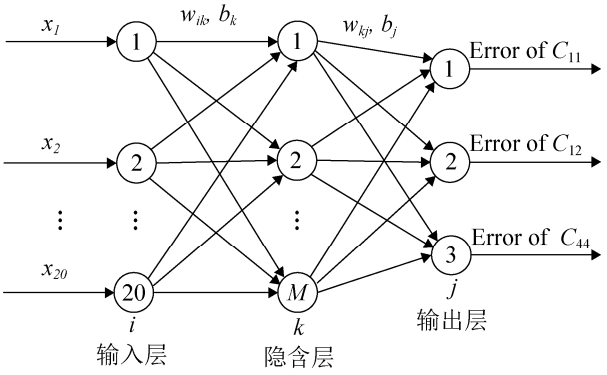


图 1 二元合金弹性常数 DFT 计算误差估计神经网络模型
Fig. 1 The neural network structure of estimation DFT calculation error of binary alloys

给定二元合金 A_xB_y，本文用以下 20 个变量作为模型输入变量，定量估计 DFT 计算模拟出的 A_xB_y 的弹性常数误差。输入变量类别为：单质和二元合金晶体结构(crystal system)相似度；原子质量(atomic mass)，两种单质元素的原子质量；原子浓度(atomic concentration)，二元合金结构 A_xB_y 中元素 A 和 B 的浓度，$c_A = x/(x+y)$，$c_B = 1 - c_A$；原子总数(total number of atoms)，用于 DFT 计算的二元合金结构和单质结构中的原子总数；键长(bond length)，二元合金结构和单质结构中两个原子间的

最短距离；单质体积，DFT 计算出的单质 *A* 和单质 *B* 的体积；单质单点能，DFT 计算出的单质 *A* 和单质 *B* 的能量；单质弹性常数计算误差，与二元合金 A_xB_y 弹性常数 DFT 计算模拟主要输入参数取值相同条件下，单质 *A* 和单质 *B* 的弹性常数 DFT 计算模拟数据误差。

对于一个特定的二元合金结构，弹性常数 DFT 计算误差估计模型输入参数中，原子质量, 原子浓度, 以及原子总数是固定的常数，不会引起模型的摄动性；键长为两类原子间的最短距离，通过材料结构计算得到，其值也是固定的。二元合金两种组成元素构成的单质的计算模拟体积、能量和弹性常数计算误差反映了 DFT 计算输入参数设置的合理性，其一定程度上也反映出不同 DFT 计算输入参数组合下，二元合金弹性常数计算模拟值误差的变化趋势。组成二元合金的两种元素单质和二元合金的晶体结构相似度反映了化合物和单质的结构差异，其值可看作模型的修正因子，对模型在不同晶体结构类型材料上的适用性差异进行修正。实际的应用中，晶体结构相似度很难准确地计算出来，往往根据不同的应用需求进行定义。晶体结构相似度的定义粒度不同，计算方法不同，其值也不同，一定程度上会引起模型的摄动性。

3 模型训练与仿真分析

3.1 数据集及模型训练参数选取

我们从文献中搜集到了 63 种单质和 45 种立方晶系二元合金的弹性常数实验数据，然后用 DFT 计算模拟软件包计算这 108 种材料的弹性常数。DFT 计算的势函数用 PAW(projector-augmented wave)，交换关联函数选用 GGA 泛函类别下的 PBE 泛函、PW91 泛函和 LDA 泛函类别下的 LDA 泛函，控制平面波基组的截断能 ENCUT、FFT 格点的参数 PREC 取 High，Normal，Accurate 三个值。一共 9 组参数，对上述 108 种材料分别用 9 组 DFT 计算参数进行计算，需要计算 972 次。用高通量材料计算平台 MatCloud 计算不同参数组合下，108 种材料的弹性常数，最多只需要提交 9 次高通量作业，便可完成 972 次弹性常数模拟计算，并且弹性常数、结构信息被自动从 DFT 计算输出文件中抽取出来，基于所搜集的弹性常数实验数据集，弹性常数计算模拟误差被自动计算。972 次弹性常数 DFT 计算模拟中，有 567 次是计算单质的弹性常数及其对应误差；另外 405 次是计算二元合金的弹性常数及其对应误差。除去 DFT 计算失败的计算例子，共搜集到 381 条样本数据。DFT 计算失败的例子有两种情况，一是由于参数设置不合理导致 DFT 计算模拟软件包计算出错，纠错 5 次后仍计算失败，共 9 种计算例子；二是由于部分元素缺失相应赝势文件导致预先设定的计算参数无法计算，共 15 种计算例子，例如 Nd、Tb、Pr、Yb 四种元素没有 PAW_LDA 赝势。以 8:2 的比例将样本数据集划分为训练数据集和测试数据集，数据来源及分类如表 2 所示。

表 2 样本数据来源及训练和测试数据集

Tab. 2 The data source and data sets for training and testing

DFT 计算模拟参数组合	每组参数计算二元合金种类数	样本数目	
		训练样本	测试样本
PT=PAW_PBE,PREC=accurate	43		
PT=PAW_PBE,PREC=high	43		
PT=PAW_PBE,PREC=normal	43		
PT=PAW_GGA,PREC=accurate	44		
PT=PAW_ GGA,PREC=high	45	305	76
PT=PAW_ GGA,PREC=normal	45		
PT=PAW_LDA,PREC=accurate	39		
PT=PAW_ LDA,PREC=high	40		
PT=PAW_ LDA,PREC=normal	39		

PT: PseudopotentialType 赝势类型

以表 2 中的 305 条样本数据为训练样本，用 tsPSO-BP 算法对图 1 中的二元合金弹性常数 DFT 计算误差估计神经网络模型进行训练。并比较 BP

神经网络训练方法、tsPSO-BP 神经网络训练方法和用于回归的支持向量机 (Support Vector Machine for Regression, SVR) 方法性能。因为弹性常数计算模拟数据有可能大于实验数据，也有可能小于实验数据，并且对于不同材料，用相同的 DFT 计算输入参数计算出的弹性常数误差相差较大，因此样本数据的输入和输出一律归一化到[–1，1]范围内。

实验过程中，3 种方法采用的参数如下。对于 SLFN 神经网络模型，我们选择具有代表性的 Levenberg-Marquardt 学习算法进行实验，神经元激活函数选用“tansig”函数：$f(x)=\exp^{x}-\exp^{-x}/\exp^{x}+\exp^{-x}$。神经网络初始隐藏层神经元个数取 4，每次递增 1，最大神经元个数设为 35。基于 7 折交叉验证方法选择接近的隐含层神经元数。tsPSO-BP 方法中，动量惯性系数采用文献[11]中提到的线性递减法进行调整，粒子数为 20，粒子进化代数设为 50; c_1=c_2=2; T_b=3; T_g=4。对于 SVR 方法，核函数选择径向基函数，我们使用组合方法选择最优参数 C 和 γ：C=[$2^{10,}$ 2^{9}，…，2^{-10}]，γ=[2^{10}，2^{9}，…，2^{-10}]。共有 20×20=400 种组合，对每一种组合，进行 7 折交叉验证实验选择最优的 C 和 γ，利用最优的 C 和 γ，对模型进行最后的训练，得到最终的 SVR 二元合金材料弹性常数 DFT 计算数据误差预测模型。

3.2 模型训练和测试性能分析

因为对于不同的材料和不同计算输入参数组合，弹性常数计算模拟数据误差相差较大，所以本文用均方差(Root Mean Square Error，RMSE)和平均相对绝对误差(Relative Absolute Error，RAE)两种性能度量方法作为二元合金弹性常数计算模拟数据误差预测模型性能的评价标准。BP、SVR 和 tsPSO-BP 3 种方法在表 2 数据集上的训练和测试性能如表 3 所示。

表 3 中汇总的实验结果均是每种算法在数据集上训练达到的最优的实验结果，每种算法重复训练 10 次，取其平均值作为最后结果。从表 3 可以看出，BP 和 tsPSO-BP 神经网络训练算法在测试数据集上的测试 RAE 比 SVR 要小，说明神经网络方法更能扑捉到单质结构、二元合金结构、单质弹性常数 DFT 计算误差和二元合金弹性常数 DFT 计算误差之间的关系，具有较好的泛化性能(generalization performance)；tsPSO-BP 比 BP 的训练和测试 RAE 和 RMSE 均小，说明 tsPSO-BP 神经网络训练方法比单纯用 BP 算法训练神经网络的优化性能要好。在测试数据集上，tsPSO-BP 方法预测的 C_{11}，C_{12} 和 C_{44} 的 DFT 计算误差的准确率分别达到 88.19%，87.83%和 88.26%，比 BP 预测的准确率分别提高了 2.06%，2.14%和 1.33%，比 SVR 方法分别提高了 5.26%，3.19%和 5.49%。

表 4 汇总了 3 种算法在训练阶段的时间消耗，从表 4 可以看出，BP 算法训练时间最短，SVR 的训练时间最长，比 BP 算法慢大约 50 倍，tsPSO-BP 的训练速度比 BP 训练速度慢 3 倍左右。从表 3 和表 4 看出，BP 的时间性能和泛化性能均高于 SVR 方法，由于粒子迭代需要时间，因此 tsPSO-BP 的训练时间要比 BP 的训练时间稍长，但精度比 BP 高。

表 3 三种不同算法的 RMSE 和 RAE 比较
Tab. 3 RMSE and RAE errors for training and testing data in BP, tsPSO-BP and SVR

	BP 的 RMSE		BP 的 RAE		tsPSO-BP 的 RMSE		tsPSO-BP 的 RAE		SVR 的 RMSE		SVR 的 RAE	
	训练集	测试集	训练集	测试集	训练集	测试集	训练集	测试集	训练集	测试集	训练集	测试集
C_{44}	0.065	0.071	12.96%	19.49%	0.052	0.062	11.74%	18.16%	0.072	0.084	14.42%	19.91%
C_{11}	0.073	0.081	13.78%	19.23%	0.062	0.071	11.81%	17.17%	0.081	0.098	15.85%	21.11%
C_{12}	0.076	0.082	14.08%	17.82%	0.068	0.072	12.17%	15.68%	0.078	0.095	16.12%	19.31%

表 4 三种不同算法的时间比较
Tab. 4 The comparison of training time for different algorithms

方法	时间/s		
	C_{44}	C_{11}	C_{12}
tsPSO-BP	36.98	34.38	33.45
SVR	595.74	776.11	599.45
BP	10.06	12.49	11.77

为进一步考察模型输入中晶体结构相似度对模型摄动性的影响，我们首先粗粒度定义相似度，当单质和二元和合金晶体结构相同时，相似度取值取为 1，否则取 0. 之后，较细粒度定义了晶体结构相似度，计算公式为：

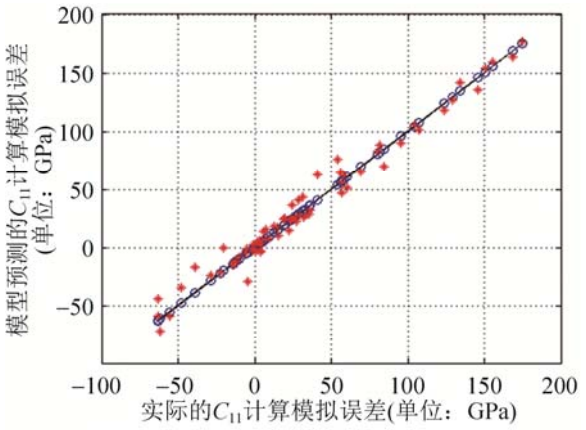
$$\mathrm{sim} = N_h / N_l \tag{9}$$

式中，N_h 和 N_l 分别为高对称性和低对称性结构的独立弹性分量个数。比如，六角晶系 Zn 的独立弹性常数分量为 $C_{11}, C_{12}, C_{13}, C_{33}, C_{44}$ 共 5 个，立方晶系 AuZn 的独立弹性常数分量为 C_{11}, C_{12}, C_{44} 共 3 个，那么 Zn 和 AuZn 的晶体结构相似度为 0.6。相较于粗粒度定义的晶体结构相似度，细粒度定义晶体结构相似度后，模型预测的 C_{11}, C_{12} 和 C_{14} 的 DFT 计算误差平均精度提高了 2.09%，1.1%，1.86%。可以看出合理地定义晶体结构相似度，会提高模型估计的准确性。

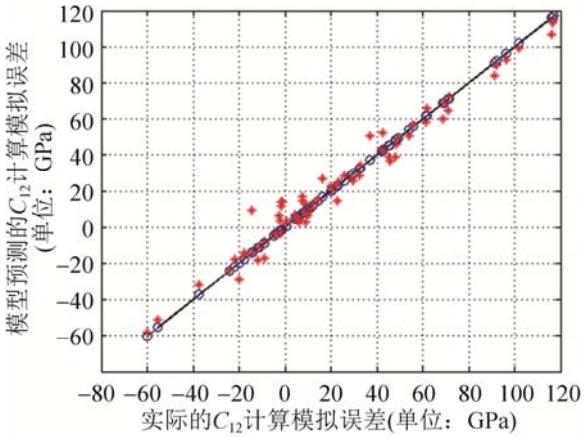
3.3 模型仿真应用

为了检验模型的应用性能，我们用未参与模型训练的测试数据集对用 tsPSO-BP 方法训练的神经网络模型进行仿真应用。图 2 给出了模型估计出的弹性常数计算模拟误差和实际的弹性常数计算模拟误差的比较，其中红色“*”表示模型的预测值，蓝色的“o”表示实际的弹性常数 DFT 计算误差。斜线上的点表示模型预测出的弹性常数计算模拟误差和实际的弹性常数计算模拟误差相等。

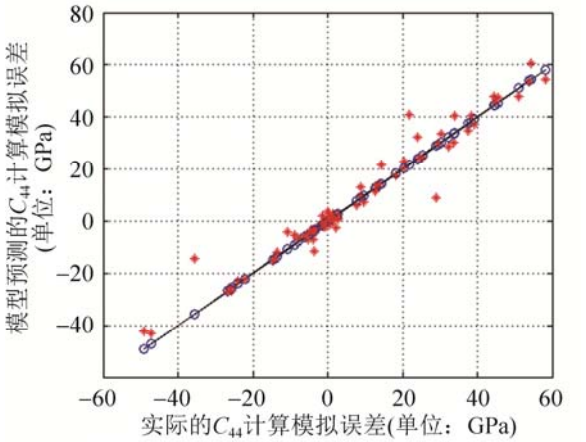
从图 2 可以看出，模型预测出的计算模拟数据误差分散在斜线周围，约 90%的预测值落在斜线上，和实际的计算模拟误差吻合的较好。从图 2 还可以看出，3 个弹性分量 C_{11}，C_{12} 和 C_{44} 的 DFT 计算模拟数据误差大部分散布在 0GPa 周围，说明 DFT 计算模拟软件包计算出的弹性常数数据具有可参考性。



(a) 测试数据集上C_{11}计算模拟误差模型预测值和实际值比较



(b) 测试数据集上C_{12}计算模拟误差模型预测值和实际值比较



(c) 测试数据集上C_{44}计算模拟误差模型预测值和实际值比较

图 2 模型仿真结果和实际值的比较
Fig. 2 Relationship of predicted values and experimental values

DOI: 10.16182/j.issn1004731x.joss.201810030

最后，我们以 $GeNi_3$，$MgCu_2$，Ni_3Fe，Cu_3Au，FeAl，Al_2La，Au_3Ag 7 个二元合金为测试材料，用训练后的模型预测 DFT 计算模拟参数取不同值时，DFT 计算出的各弹性常数分量的误差。我们将弹性常数 DFT 计算模拟值和实验值分别记为 V_{DFT} 和 V_{Exp}。用两种数学形式表示误差，一种是 V_{DFT} 和 V_{Exp} 的差值，用 $E_{DFT}=V_{DFT}-V_{Exp}$ 表示，另一种是相对误差，表示为 $R_{DFT}=(E_{DFT}/V_{Exp})\times 100\%$。模型仿真结果如表 5 所示，由于篇幅的限制，表 5 仅列出了部分 DFT 计算参数下模型预测的 DFT 计算数据误差和实际的误差，以及模型误差。

表 5 中的第 3 列为模型估计的 7 种测试材料在 DFT 计算参数取不同值时弹性常数 C_{11}，C_{12} 和 C_{44} 的 DFT 计算模拟误差，第 4 列为实际的 DFT 计算模拟误差，第 5 列为模型误差，模型误差为模型估计的 DFT 计算模拟误差和实际的 DFT 计算模拟误差的差值。从表 5 可以看出，模型的仿真误差大多在[-5Gpa，5Gpa] 范围内，表明本文提出的模型具有较好的准确性，具有可用性。

目前，材料计算模拟者通常依靠经验或者多次尝试计算寻找适合所计算体系的参数进行 DFT 计算模拟，以获得较可靠的计算模拟值，浪费时间和计算资源。本文所建立的二元合金弹性常数 DFT 计算模拟数据误差估计模型不需要进行弹性常数计算，便可估计出不同参数组合下弹性常数 DFT 计算误差，从而选择误差较低的参数进行计算，节约了计算成本。

表 5 模型在测试材料上的仿真应用结果

Tab.5 The model prediction results for testing materials

二元合金	DFT 计算参数组合	模型预测的 E_{DFT} (Unit: GPa) R_E			实际的 E_{DFT} (Unit: GPa) R_E			模型误差 (Unit: GPa)		
		C_{11}	C_{12}	C_{44}	C_{11}	C_{12}	C_{44}	C_{11}	C_{12}	C_{44}
$GeNi_3$	XC=PBE PREC=accurate	81.04 30.81%	21.27 14.88%	29.43 28.57%	80.35 30.55%	22.27 15.58%	33.62 32.65%	0.69	–1	–4.19
$GeNi_3$	XC=LDA PREC=normal	148.41 56.43%	70.41 49.24%	56.35 54.71%	150.76 57.32%	68.95 48.22%	58.28 56.58%	–2.35	1.46	–1.93
$MgCu_2$	XC=PBE PREC=high	14.54 11.63%	23.94 33.39%	1.20 2.84%	19.97 15.97%	16.50 23.01%	–0.04 –0.09%	–5.43	7.44	1.24
$MgCu_2$	XC=GGA PREC=normal	96.51 77.21%	56.78 79.20%	17.14 40.51%	95.78 76.63%	55.72 77.72%	18.29 43.25%	0.73	1.06	–1.15
Ni_3Fe	XC=PBE PREC=normal	168.58 73.16%	64.89 44.93%	48.20 40.44%	168.66 73.20%	69.11 47.86%	51.04 42.82%	–0.08	–4.22	–2.84
Cu_3Au	XC=PBE PREC=normal	57.14 30.54%	9.96 7.38%	20.92 30.86%	56.11 29.99%	7.71 5.72%	20.36 30.04%	1.02	2.25	0.56
FeAl	XC=LDA PREC=high	130.22 71.92%	39.29 34.56%	42.74 33.63%	134.29 74.17%	43.09 37.90%	34.00 26.75%	–4.06	–3.8	8.74
Al_2La	XC=GGA PREC=normal	–13.07 –8.82%	–0.72 –2.25%	0.45 1.03%	–13.45 –9.08%	–1.91 –5.99%	–0.83 –1.89%	0.38	1.19	1.28
Al_2La	XC=LDA PREC=high	–1.27 –0.85%	4.11 12.89%	1.18 2.71%	3.58 2.42%	5.30 16.62%	1.46 3.35%	–4.85	–1.19	–0.28
Au_3Ag	XC=PBE PREC=normal	5.04 3.02%	–11.94 –9.01%	–3.67 –7.54%	4.95 2.98%	–11.2 –8.45%	–3.25 –6.68%	0.08	–0.74	–0.42
Au_3Ag	XC=GGA PREC=accurate	3.89 2.33%	–11.67 –8.81%	–3.38 –6.96%	2.07 1.24%	–13.85 –10.45%	–4.33 –8.92%	1.82	2.18	0.95

4 结论

本文针对材料计算模拟数据误差估计的问题进行分析，以立方晶系二元合金的弹性常数性质为例，建立误差预测神经网络模型，估计不同计算输入参数组合下弹性常数计算模拟结果的误差，使得不需要进行 DFT 计算模拟即可了解到利用不同输入参数组合进行 DFT 计算模拟得到的数据的误差，从而选择对应计算模拟结果误差较低的参数进行计算，提高 DFT 计算模拟数据质量，增强计算模拟数据的可用性。通过分析传统 BP 算法和 PSO 算法各自的优缺点，结合二者优势，将简易 PSO 算法引入到 BP 神经网络训练中，在粒子进化后期，对个体极值增加随机扰动，将全体极值替换为 BP 算法训练后的 SLFN 模型连接权值，并将适应度最差粒子的个体极值替换为历史全局极值，增强粒子对历史的记忆能力。将改进后的 tsPSO-BP 算法应用于立方晶系二元合金的弹性常数计算模拟数据误差估计中。仿真结果表明，相比较于 SVR 和 BP 算法，tsPSO-BP 算法的泛化性能较高，用其训练的模型的预测值和实际值吻合的较好。在测试数据集上，tsPSO-BP 方法预测的 C_{11}，C_{12} 和 C_{44} 的 DFT 计算误差的准确率分别达到 88.19%，87.83%和 88.26%。由模型仿真结果可以看出，模型具有较高的预测准确率。在下一步的研究中，我们计划用人工智能方法对利用最优输入参数计算模拟得到的结果进行修正，使修正后的数据更可靠，更好的应用于新材料的设计中。

参考文献:

[1] National Science and Technology. Materials Genome Initiative for Global Competitiveness [R]. Washington DC, USA: Office of Science and Technology Policy, 2011.

[2] Hohenberg P, Kohn W. Inhomogeneous Electron Gas [J]. Physical Review (S1550-2376), 1964, 136(3B): B864-B871.

[3] Parr R G, Yang W. Density-Functional Theory of atoms and molecules [M]. New York, USA: Oxford University Press, 1989.

[4] 王娟, 杨小渝, 王宗国, 等. SQS 二元合金设计的高通量方法和技术研究 [J]. 计算机工程与科学, 2016, 38(255): 401-410.
Wang Juan, Yang Xiaoyu, Wang Zongguo, et al. A high-throughput Computational Approach for SQS Binary Alloy Design [J]. Computer Engineering & Science, 2016, 38(3):401-410.

[5] 王卓, 杨小渝, 郑宇飞, 等. 材料基因组框架下的材料集成设计及信息平台初探 [J]. 科学通报, 2013, 58(35): 3733-3742.
Wang Zhuo, Yang Xiaoyu, Zheng Yufei, et al. Integrated Materials Design and Informatics Platform within the Materials Genome Framework [J]. Chin Sci Bull (Chin Ver), 2013, 58: 3733–3742.

[6] Xiujung Wang, Lihong Hu, Laiho Wong, et al. A Combined First-principles Calculation and Neural Networks Correction Approach for Evaluating Gibbs Energy of Formation [J]. Molecular Simulation (S0892-7022), 2004, 30(1): 9-15.

[7] Ozerdem M S, Kolukisa S. Artificial Neural Network Approach to Predict the Mechanical Properties of Cu–Sn–Pb–Zn–Ni cast alloys [J]. Materials & Design (S0264-1275), 2009, 30(3): 764-769.

[8] 邓万宇, 郑庆华, 陈琳, 等. 神经网络极速学习方法研究 [J]. 计算机学报, 2010, 33(2): 279-287.
Deng Wanyu, Zheng Qinglin, Chen Lin, et al. Research on Extreme Learning of Neural Networks [J]. Chinese Journal of Computers, 2010, 33(2): 279-287.

[9] Rumelhart D E, Hinton G E, Williams R J. Learning Representations by Back-propagating Errors [M]. USA: MIT Press, 1986: 533-536.

[10] Kennedy J, Eberhart R C. Particle Swarm Optimization [C]// Proc. of the IEEE Conf. on Neural Networks, IV. Perth, Australia: IEEE Press, 1995: 1942-1948.

[11] Shi Y, Eberhart R C. A Modified Particle Swarm Optimizer [C]// Proc of the 1998 IEEE ECP. Piscataway, NJ, USA: IEEE, 1998: 69-73.

[12] 胡旺, 李志蜀. 一种更简化而高效的粒子群优化算法 [J]. 软件学报, 2007, 18(4): 861-868.
Hu Wang, Li Zhishu. A Simpler and More Effective Particle Swarm Optimization Algorithm [J]. Journal of Software, 2007, 18(4): 861-868.

[13] 张捍东, 陶刘送. 粒子群优化 BP 算法在液压系统故障诊断中应用 [J]. 系统仿真学报, 2016, 28(5): 1186-1190.
Zhang Handong, Tao Liusong. Application of PSO-BP Algorithm in Hydraulic System Fault Diagnosis [J]. Journal of System Simulation (S1004-731X), 2016, 28(5): 1186-1190.

(下转第 3860 页)

DOI: 10.16182/j.issn1004731x.joss.201810030