应用机器学习加速新材料的研发

吴炜^{1,2}, 孙强^{1,2*}

1. 北京大学工学院材料系, 北京 100871;

2. 北京大学应用物理与技术研究中心, 北京 100871

*联系人, E-mail: sunqiang@pku.edu.cn

收稿日期: 2018-03-17; 接受日期: 2018-04-23; 网络出版日期: 2018-07-31

国家自然科学基金(编号: 21573008, 21773003)和国家重点研发项目(编号: 2016YFB0100200, 2017YFA0204902)资助

摘要 材料不仅是国民经济的基础, 而且也是高新技术的载体. 超越常规手段、应用新方法加速新材料的研发已成为全世界的研究热点. 随着数据驱动方法取得的巨大成功, 机器学习受到了日益高度的关注. 它结合计算机科学、数据库理论、统计学、计算数学和工程学, 不仅能展现出更快的计算速度和可靠的预测能力, 大幅度提升材料计算效率, 而且还能有效地处理一些难以运用传统模拟计算方法解决的体系和问题, 这为研发具有特殊功能和特殊结构的新材料以满足日益提升的新技术的要求提供了契机. 本文将简要概述机器学习的基本原理, 介绍机器学习模型中的几种典型算法以及机器学习在新材料研究中的应用进展, 并对机器学习在材料科学领域中的未来的发展前景做出展望.

关键词 新材料, 机器学习, 材料计算模拟

PACS: 02.70.-c, 61.50.Ah, 71.15.-m

1 引言

伴随着计算机和网络的飞速发展, 如何使用计算机实现智能化已成为了一个备受关注的课题. 虽然人工智能的研究兴起已久, 但是机器学习仍然是人工智能领域中较为年轻的一个分支. 近年来, 机器学习受到了人们的广泛关注, 尤其是谷歌开发的人工智能 AlphaGo 在与围棋世界冠军的人机大战中频频取胜^[1,2], 将机器学习的巨大潜力展现在人们的视野当中. 机器学习已被广泛应用于计算机视觉^[3]、自然语言处理^[4,5]、数据挖掘^[6]、机器人应用^[7,8]、战略类游

戏^[1,2]、医学诊断^[9,10]、证券市场分析^[11]等, 并取得了巨大的成功. 另一方面因为材料科学领域在信息时代中产生的大数据, 机器学习方法在材料科学领域的应用便应运而生^[12,13]. 在材料计算领域, 目前应用最为广泛的当属密度泛函理论. 然而, 目前的理论计算结果与实际应用之间仍然存在一定的差距, 例如, 密度泛函理论在高温、高压以及强磁场的环境下模拟计算的结果就不尽人意. 运用机器学习方法可以大大节省计算时间, 从而应用于更加复杂的体系或者更长的时间尺度, 因此, 将机器学习与现有的计算方法相结合, 便会使得新材料设计的效率更高、准确度更好, 从而为实验提

引用格式: 吴炜, 孙强. 应用机器学习加速新材料的研发. 中国科学: 物理学 力学 天文学, 2018, 48: 107001

Wu W, Sun Q. Applying machine learning to accelerate new materials development (in Chinese). Sci Sin-Phys Mech Astron, 2018, 48: 107001, doi: 10.1360/SSPMA2018-00073

供更加有效的指导.

2 机器学习的原理及分类

机器学习^[14,15]是一门交叉学科, 它将计算机科学、统计学、数学与工程学的相关知识结合起来形成人工智能的一个重要分支, 它所关注的问题是: 计算机程序如何随着经验积累自动提高性能^[16]. 换言之, 机器学习就是计算机利用已知数据, 在某些算法的指导下自动优化并改进模型, 使之能对全新的情境进行判断和预测. 实际上机器学习的思想来源于人类对智能学习的思考与认知. 学习能力是人类智能的一种体现, 人类的学习过程, 都需要经过人类的大脑完成. 大脑就像一台紧密的仪器, 与外界环境进行信息交互, 总结规律, 形成经验, 在遇到新问题时形成及时的判断与决策. 概括来说, 人类的学习过程都要经历经验累积、规律总结, 最终达到灵活运用于解决问题的阶段. 类似的, 机器学习可以被分为输入、学习、输出3个阶段. 输入阶段计算机识别并储存大量的训练数据, 在学习阶段寻找数据中的内在联系, 输出阶段通过学习到的规律对新的问题进行判断和预测. 机器学习与人类的经验学习过程趋于相同, 就是计算机模仿人类理解、思考和创造的过程. 机器学习的流程如图1所示, 首先是数据的输入阶段, 包括了数据的收集以及预处理; 然后是训练学习阶段, 通过训练数据对模型不断更新, 并用更新后的模型对测试数据做出预测和对比, 根据结果进一步更新模型直至达到期望的优化目标; 最后的输出阶段就是利用优化好的模型对于未知的数据做出预测或者决策.

人类的学习过程是依赖于过往的经验, 同样的, 机器学习也是依赖于给定的数据的. 从本质上讲, 机器学习是一门数据驱动的方法, 数据是机器学习的基础, 它既可以作为系统学习的知识来源, 也可以作为评判

算法优劣的标准. 数据的真实性和可靠性决定了机器学习结论的有效性, 这也是机器学习与大数据密切相关的重要原因. 根据所处理的数据类型的区别, 机器学习可以被分为监督式学习、非监督式学习、半监督式学习和强化学习. 其中监督式学习是指训练数据如 (x, y) 数据对, 目标是在通过训练数据优化算法后, 查询输入为 x^* 时产生一个预期输出 y^* . 相对应的, 如果训练数据中 y 为未知的标签, 也就是说对于训练数据的每一个 x , 其对应的预期输出 y 都是未知的, 那么这一类机器学习的方法叫作非监督式学习, 其目的是寻找和挖掘数据中有意义的模式, 如聚类、关联规则挖掘等等. 而半监督式学习则介于监督式学习和非监督式学习之间, 数据中包含一部分成对的输入输出数据对和一部分没有相对应的预期输出的数据; 强化学习则是对行为和表现给予奖励和惩罚信号, 并以此来学习如何在环境中获得更好的表现.

从数学的角度来看, 机器学习由数据和方法构成, 方法包含策略、模型和算法3个要素. 机器学习的过程, 就是选取合适的方法, 对数据进行处理, 从而完成学习的目的. 以监督式学习的方法为例, 模型是需要学习的条件概率分布或者决策函数; 策略就是学习准则, 确定学习准则来选取最优模型; 算法则是具体的求解最优模型的计算方法. 学习策略是度量模型好坏的标准, 模型的好坏, 包括使用模型进行预测和决策的准确度以及模型本身的复杂度. 为了度量模型在预测时的错误程度, 引入了损失函数, 损失函数是用来度量对于给定的输入 x , 由模型得到的预期输出 $f(x)$ 与真实值 y 之间的差异, 常用的损失函数包括0-1损失函数、平方损失函数、对数损失函数等等. 在数据集较大, 样本容量足够大的情况下, 得到的损失函数的期望值越小, 得到的模型就越好, 但是, 当样本容量比较小时, 损失函数的值小不一定代表模型好, 还有可能出现过拟合的现象. 过拟合现象的出现, 代表着优化得到模型过于复杂, 虽然对给定的训练数据的拟合结果很好, 但是模型的泛化能力不足. 此时就需要考虑模型的复杂度, 引入一个包含复杂度的惩罚项, 权衡损失函数和惩罚项, 优化得到综合性能最好的结构风险最小化模型.

图 1 (网络版彩色)机器学习的三个阶段

Figure 1 (Color online) Three stages of machine learning.

3 机器学习的算法

算法是机器学习中学习模型的具体计算方法. 机

机器学习基于训练数据集, 根据学习策略, 从模型的假设空间中选择最优模型, 这时候机器学习的主要问题就变成了求解最优化问题. 因为通常情况下这个最优问题的解析解是不存在的, 所以需要数值方法进行计算求解, 如何找到全局最优解, 并且保证求解过程的高效性, 是机器学习的算法要解决的重要问题.

机器学习的算法种类繁多, 机器学习的不同功能以及对应的一些算法都在图2中列举出来. 一些常见的算法^[17]包括朴素贝叶斯、k近邻、决策树、支持向量机、提升算法、EM算法、人工神经网络、k均值聚类、Apriori算法等等, 有些算法发展较为成熟, 有些算法诞生不久, 但是各种算法都各有千秋, 本文中简单介绍几种影响力比较大的传统算法.

3.1 决策树

决策树^[18,19]是一种基本的分类与回归算法. 通常情况下, 决策树模型是一种利用树形结构对样本进行分类的模型算法, 主要算法包括ID3算法和C4.5算法, 此外, CART算法是既可以用于分类也可以用于回归, 这3种算法构成了决策树模型的主要内容.

一个决策树的构造如图3所示, 决策树由节点和有向边组成, 节点包括内部节点和叶节点, 内部节点代表一个特征, 叶节点表示一个类别. 决策树就是选择一定的特征对数据进行判别, 根据判别的结果将数据划分到不同的子节点中, 每一个子节点代表该特征的一个取值, 不断选择特征进行分类, 直至达到叶节点, 也就是所有的数据被分到相应的类别之中. 决策树表示的是在给定特征条件下不同类别的条件概率分布, 决策

图3 (网络版彩色)决策树的示意图. 图中的决策树模型为一个简单的二叉树模型, 圆圈代表内部节点, 方框代表叶节点, 内部节点代表用于分类的一个特征, 叶节点代表一个类别

Figure 3 (Color online) A schematic diagram of a decision tree. The decision tree model in the figure is a simple binary tree model. Circles are for internal nodes, boxes for leaf nodes, the internal nodes represent a feature for classification, and leaf nodes for a category.

树实质上就是从训练数据中归纳出一组分类的规则, 不同的决策树代表了复杂程度不同的分类模型. 对于给定的训练数据集, 可能有多个与之不发生矛盾的决策树模型, 也可能没有符合条件的决策树模型, 应用过程中需要根据实际情况, 选择对训练数据拟合结果比较好, 泛化能力强, 对未知数据有良好预测能力的模型.

决策树算法主要包括特征选择、决策树的生成和剪枝3个步骤. 特征选择主要是选取对训练数据具有优良分类能力的特征, 选择合适的特征进行分类能有效提高决策树的学习效率. 具体的特征选择的标准通常是信息增益或者信息增益比. 信息增益是表征某个特征对于整个数据集的分类的不确定性减少的程度, 不同的特征往往具有不同的信息增益, 信息增益越大, 该特征的分类能力越强. 在确定了特征选择的准则后, 就需要确定决策树的生成和剪枝. 对于决策树的生成, ID3算法和C4.5算法分别以信息增益和信息增益比作为特征选择的准则, 选择合适的特征, 将数据集分割为若干非空子集, 构成决策树的分支和子树, 直至分类完成. 对于决策树的剪枝, 则是基于决策树模型的损失函数最小化的思想进行的, 决策树的生成只考虑了提高信息增益或者信息增益比来提升对数据的拟合程度, 而决策树的剪枝则是通过优化损失函数, 权衡模型复杂度和准确率, 达到模型的结构风险最小化. 由于决

图2 (网络版彩色)机器学习的主要功能和相应的算法

Figure 2 (Color online) Main functions and corresponding algorithms of machine learning.

策树表示一个条件概率分布, 所以, 决策树的生成本质上就是模型的局部最优化选择, 决策树的剪枝则是考虑的全局最优化.

3.2 朴素贝叶斯

朴素贝叶斯算法^[20]是建立在贝叶斯定理上的一种算法, 可以用于分类的计算. 贝叶斯定理给出了条件概率的推理算法. 朴素贝叶斯算法的基本思想是, 对于给出的待分类项 x , 求解这个待分类项条件下的各个类别 y_i 的出现概率, 待分类项的类别就是条件概率最大的那个类别. 朴素贝叶斯算法的最大特点就是认为待分类项的各个特征之间都是相互独立的, 这也是朴素贝叶斯算法应用的一个重要前提.

由朴素贝叶斯算法处理的分类流程如图4所示, 在准备阶段中需要确定特征属性和获取训练数据, 在训练阶段根据分类器的算法进行计算, 得到各个类别的出现概率和各个类别下出现各个特征属性的条件概率, 应用阶段就是对待分类项进行计算处理, 得到待分类项与类别的映射关系.

朴素贝叶斯算法的优点是训练时间短, 并且这种算法是基于坚实的数学基础, 有着稳定的分类效率. 朴素贝叶斯算法的缺点就是在于这种算法的前提在实际问题中难以满足, 朴素贝叶斯算法的前提是假设各个特征之间相互独立, 这在实际问题尤其是复杂问题

图 4 (网络版彩色)朴素贝叶斯算法的流程图

Figure 4 (Color online) Flow chart of Naïve Bayes algorithm.

中往往是无法满足的, 所以朴素贝叶斯算法在处理实际问题时, 可能准确度并不能达到理想的高度.

3.3 支持向量机

支持向量机^[21]是一种二分类模型, 其学习策略是间隔最大化, 通过构造一个超平面将多维空间分为两个区域, 不同的输入 (x, y) 落在不同的区域里, 因此可以应用于如图5所示的元线性分类. 对于二维空间下的分类问题, 分类超平面就是一条直线, 圆圈和叉分别代表正分类点和负分类点. 推算到 n 维空间中 $(n \geq 3)$, 即采用一个 $n-1$ 维的超平面将正分类点与负分类点分离.

根据训练数据的性质, 支持向量机可以被分为3类: 当训练数据线性可分时, 通过硬间隔最大化, 学习一个线性分类器, 即线性可分支持向量机; 当训练数据近似线性可分时, 通过软间隔最大化, 也学习一个线性分类器, 即线性支持向量机; 当训练数据线性不可分时, 通过使用核技巧, 将低维度的非线性问题转化为高维度下的线性问题, 学习得到非线性支持向量机.

支持向量机的基本思想是求解能正确分类训练数据, 得到使几何间隔最大的超平面. 以最简单的线性可分问题为例, 对于线性可分的数据集, 能够正确分离数据的超平面有无穷多个, 然而几何间隔最大的超平面是唯一的. 几何间隔最大, 说明这个模型可以将数据集中最难分的点以最大的确信度分离开, 也就是说几何间隔最大的超平面能以足够大的确信度对所有训练数据进行分类, 这样的超平面在应对未知数据时也应该有足够好的分类能力. 在确定几何间隔最大的超平面的过程中, 只有与超平面最近的样本点起到作用,

图 5 (网络版彩色)线性二分类问题示意图

Figure 5 (Color online) A schematic diagram of linear Dichotomous problem.

这样的样本点被叫作支持向量, 这种分类模型也因此被称作支持向量机。

在实际问题中, 数据在多维空间内通常并非是线性可分的, 也就是说在数据所在的输入空间内, 不存在一个超平面能够完成要求的分类. 一个可行的解决方法是应用核技巧^[22], 将数据从输入空间通过一个特定的函数映射到更高维度的空间内, 并且在更高维度的空间内寻找超平面, 我们将这样的更高维度的空间称为特征空间. 因为将输入空间通过一个特定的映射到更高维度的特征空间之中, 所以在高维度的空间内的计算量会显著增大, 计算复杂度也大大提高. 为了减少计算量, 在假设支持向量机的计算只涉及内积计算的前提下, 引入一个核函数, 将高维度的特征空间中的内积计算转化为低维度的输入空间中的数据 (x, y) 的内积运算的非线性变换. 有时候非线性映射函数过于复杂无法直接表示出来, 而核函数相对比较容易寻找, 所以寻找合适的核函数对于支持向量机起到至关重要的作用. 常用的核函数包括线性核函数、多项式核函数和高斯核函数等, 采用不同的核函数将导致不同的支持向量机算法。

支持向量机在处理线性回归问题以及二元分类问题上都非常有效, 但是因为支持向量机自身算法上的特点与限制, 对于多元分类问题无法单独使用单一的支持向量机解决。

3.4 频繁模式挖掘

频繁规则挖掘^[23]属于非监督式学习, 主要用于挖掘和发现大量数据中项集之间的有趣的关联. 频繁规则挖掘的一个典型案例就是购物篮分析, 即通过挖掘顾客购买的不同商品之间的联系, 分析顾客的购买习惯, 通过了解哪些商品被顾客频繁地同时购买, 帮助制定营销策略。

关联规则的度量主要包括两个指标: 支持度和置信度. 支持度表示的是数据项集中同时包括两个事务数据A和B的概率, 对于给定的数据集, 给出预设的最小支持度, 如果数据项集满足最小支持度, 则称之为频繁项集. 置信度表示的是在包含A的条件下, 同时也包含B的条件概率, 同时满足最小支持度和最小置信度的规则称为强规则. 关联规则挖掘的过程可以分为两步, 首先找出数据项集中的频繁项, 然后由频繁项集产生强关联规则。

Apriori算法和频繁模式增长是两种常见的关联规则挖掘的算法. Apriori算法使用一种逐层搜索的迭代方法, 由包含 k 个项的频繁项集搜索 $k+1$ 项集, Apriori算法的具体流程如图6所示, 在算法的第一次迭代过程中, 每个项都是1-项集的候选项, 算法通过扫描数据集的所有事务数据, 对每个项进行计数, 根据预先设定的最小支持度推算出相应的支持计数, 从而确定包含1个项的频繁项集, 即为1-频繁项集; 在算法的第二次迭代过程中, 首先由1-频繁项集的项两两组合产生所有的二项集作为候选, 然后再扫描数据集的所有事务数据, 计算出同时包含2个项的支持计数, 并筛选得到2-频繁项集; 依此类推, 直至找出所有的频繁项集. 正是因为这样的性质, 导致Apriori算法在处理数据庞大的数据集时, 随着频繁项集中项数的增加, 对整个数据库的遍历次数也随之增加, 算法的耗时增加, 有效性降低, 另外, 长的频繁项集的备选集合的空间复杂度也会相应的水涨船高. 频繁模式增长^[24]能不通过产生候选集的情况下完成Apriori算法的功能, 它通过两次数据库的扫描就能得到项集的关联信息, 频繁模式增长的第一次扫描将数据集的1-频繁项按支持度的计数降序排列, 第二次扫描则将所有频繁项集保存到一个频繁模式树(Frequency Pattern Tree, FP-Tree)中, 同时保存关联信息. 频繁模式增长的方法将复杂的频繁

图 6 (网络版彩色)Apriori算法的流程示意图. 假设数据集中包括A, B, C, D, E 5种项, 设置的最小支持计数为2, 应用Apriori算法从1-频繁项集逐步搜索得出所有的频繁项集

Figure 6 (Color online) Flow chart of Apriori algorithm. Assuming that the data set includes five items A, B, C, D, and E with the minimum support count of 2, one can search all frequent item sets using the Apriori algorithm.

模式问题转化为递归地寻找简单的频繁模式, 由若干子树通过节点相互连接, 形成频繁模式树. 频繁模式增长能大大减小数据集的搜索开销, 在时间代价上要远胜于Apriori算法, 但是, 在数据集很大的情况下, 构建基于内存的频繁模式树也是不太现实的, 需要用到其他的替代方法, 这里就不再深入探讨.

3.5 人工神经网络

人工神经网络^[25,26]的算法以仿生学为出发点, 通过模仿人类大脑的神经元的工作方式, 将人脑的神经元抽象为一个节点, 用大量的节点将输入数据通过函数拟合与估计转化为目标的输出数据. 不同于传统的概率学推理演算, 人工神经网络通过统计学的方法, 实现类似于人的简单的判断能力和决策能力. 人工神经网络的结构如图7所示, 其节点一般分为三类: 输入节点, 接受输入的数据信息; 输出节点, 保存数据信息处理的结果; 输入节点和输出节点之间的节点成为隐藏节点, 人工神经网络的各类节点均为分层分布, 不同层数的节点之间可以由线连接, 对应于神经结构中的突触, 代表一种非线性的映射, 另外, 对于每一个映射, 还存在一个权值, 用以调整各个映射之间的权重, 进而调控隐藏节点的存储内容, 输出节点存储的数据为预期的输出, 当隐藏层数为1时, 即为简单的人工神经网络, 当隐藏层数继续增加时, 计算的复杂度也会增加, 这种运用多层隐藏层数的人工神经网络的算法也被称为深度学习^[27]. 人工神经网络的学习过程, 是通过训练数据对各个层中节点的权值进行校正而不断优化整个网络模型.

人工神经网络算法的训练过程中, 算法的优化是

图 7 (网络版彩色)人工神经网络的结构示意图

Figure 7 (Color online) Structure schematic diagram of artificial neural networks.

通过以下几个部分完成的: (1) 调整每个突触中的权值; (2) 神经网络结构的修改, 包括节点层数的增减, 层内节点数目的增减以及突触的连接与删除; (3) 非线性映射的函数的选择与更改. 但是, 因为人工神经网络的训练过程是通过减小最终的输出与预期输出之间的差别来调整输出节点之前的各个隐藏节点中的数据, 也就是说误差的来源是从输出节点倒退的向整个神经网络传播的, 这种传播方式被称为反向传播. 反向传播使得随着隐藏层数的增加, 训练阶段的算法优化越复杂, 对于误差的消除越困难, 一种迭代的方法被用于人工神经网络算法的优化之中, 用这种迭代算法使得训练数据在人工神经网络中反复迭代, 多次遍历, 以达到减少误差的目的, 但是随着隐藏层数的增加, 迭代方法的计算量也会大幅增加, 同时, 靠近输出节点的隐藏层中的节点可能发生过度拟合而影响整体的算法准确度, 所以在使用人工神经网络算法时, 对于隐藏层数的选择需要非常慎重.

4 机器学习在材料设计中的应用

在材料计算模拟中, 经常会产生大规模, 高维度的数据集, 机器学习可以提供一种可扩展的方法来识别大数据集中的模式, 提取数据中的规律和趋势, 这为材料的逆向设计提供了一种策略和思路. 目前, 机器学习已经被应用于许多材料的研究与设计^[28]之中, 包括金属有机框架材料^[29](MOFs)、软物质及生物材料^[30]、锂离子电池材料^[31,32]、热电材料^[33,34]、催化材料^[35,36]、碳材料^[37]等等. 除了可以有效加快新型材料的设计与研发过程, 机器学习还可以优化和改良现有的材料理论计算方法. 目前的理论计算方法中, 涉及的半经验参数和泛函, 大多都具有一定的特例型, 只适用一种或几种体系, 迄今为止尚且没有一种通用的系统方法, 对于这样的情况, 用机器学习去优化其中的参数是一种潜在的可行方法. 本文将选取一些有代表性的研究实例与文章, 简单介绍机器学习在材料计算领域中的应用.

4.1 预测材料的属性

目前, 材料的设计与研发主要还是依靠传统实验方法去合成与测试样品, 新材料的研发过程耗时长, 效率低. 理论上的模拟计算, 尤其是高通量筛选技术

的引入, 加快了材料的研发进程. 机器学习, 相比于现在主流的理论计算方法, 计算速度快, 灵活度高, 泛化能力强, 并且具有与实验相结合的潜力, 可以有效地加快材料的设计和研发过程^[38]. 机器学习在材料的设计与研发中的应用, 主要是借助于机器学习对于材料属性的预测能力, 这里的材料属性包括材料的组成和结构, 以及材料的微观性质与宏观性能, 还有材料的合成加工方法.

在具有特定功能的材料的研发过程中, 找到结构与性质之间的对应关系十分重要, 很多情况下, 材料的结构与性质之间并非简单的一一对应的关系, 在现有的理论知识无法解释结构与性质之间的关系时, 利用机器学习进行训练学习, 准确预测材料的结构和性质, 挖掘其中的对应关系也是一种可行的方法. 以锂离子电池的工作过程为例, 锂离子电池在工作过程中是一个动态的体系, 内在的物理化学过程也比较复杂, 锂离子电池材料的结构与性质之间的对应关系也会相应的复杂^[39], 所以对锂离子电池性质的理论模拟工作与实验数据存在一定的差距, 而机器学习方法可以直接用大量的实验数据直接学习和归纳其中的结构与性质的对应关系, 对于电池的研究工作也具有一定的指导意义. 目前机器学习已经被用于预测锂离子电池的某些属性, 在材料的结构方面, Attarian Shandiz等人^[40]应用机器学习中的不同算法对硅酸盐正极材料进行晶体结构上的分类; 在材料的微观性质方面, Fujimura等人^[41]用支持向量机算法来预测全固态电池中锂离子导体材料的离子电导率, 另外也有文章报道应用机器学习来筛选石榴石型^[42]、橄榄石型^[43,44]以及磷酸盐型^[45]的电极材料; 在材料的宏观性能方面, 机器学习也被用于研究锂离子电池的循环性能, 包括使用遗传算法^[46]、支持向量机^[47,48]、粒子滤波算法^[49]等在内的多种算法被用于评估锂离子电池的剩余有效寿命以及容量衰减情况^[50-53].

在非晶态材料的研究中, 机器学习同样取得了一定的进展. 对于玻璃态物质, 其结构与动力学性质之间只有微弱的相关性, 有实验组运用机器学习的方法, 表征玻璃态材料的局部结构的“柔软度”, 并探索了其动力学性质与“柔软度”之间的相关性^[54]. 类似的, 在无序固体材料中, 运用机器学习的方法也能有效地研究缺陷的流动与粒子的重排列^[55], 从而归纳出无序材料中具有普遍性的微观结构特征.

材料的合成与加工, 尤其是一些复杂的化合物的

反应机理, 目前尚且缺乏深刻的理解和认知, 对材料的研究与发展主要依赖于探索性的尝试与合成, 通过经验的累计来确定合理的合成与加工流程. 机器学习作为一种数据驱动的计算方法, 可以巧妙地结合实验数据, 提供了一种有潜力的工艺优化的方法. Raccuglia等人^[29]用机器学习的方法训练实验数据来预测MOFs材料合成中的钒亚硒酸盐结晶反应的实验结果, 值得一提的是, 他们用来训练支持向量机模型的训练数据, 还包括大量失败的或者不成功的实验数据, 这样的方法既可以有效节省反复实验的时间与成本, 还能充分发挥出过往不成功的实验的作用. 实验结果表明, 使用机器学习模型预测产物形成条件的成功率高达89%, 远胜过传统的人工合成策略. 另外, 在合金的加工过程中, 利用机器学习的方法也可以预测加工过程中的工艺参数与合金性能之间的关系^[56], 机器学习在这种复杂且模糊的系统下有着出色的学习和推导能力.

4.2 改进材料的理论模拟计算方法

目前应用最为广泛的理论计算方法还是密度泛函理论, 然而, 目前的理论计算结果与实际应用之间仍然存在一定的差距^[57]. 尽管如此, 用密度泛函理论求解电子结构问题依然可以提供大量有价值的结果, 并且将这些数据信息整合到大型数据库中, 如The Materials Project^[58](<http://materialsproject.org>)以及OQMD(<http://oqmd.org>)等等. 而机器学习方法在大数据的支持下可以发挥更大的作用^[59-61], 因此, 将机器学习与现有的理论计算方法结合起来, 可以加速理论模拟工作的进展.

将密度泛函理论与机器学习结合起来的方法有很多, 最简单, 最直接的方法就是利用密度泛函理论的计算结果作为机器学习的训练数据进行学习与预测^[62], 这种思想已有一些实例, Ceder实验组^[63]用机器学习和DFT方法相结合, 用于寻找自然界中不存在的三元含氧化合物. Tanaka实验组^[41]结合机器学习与第一性原理计算去设计锂离子导体材料, 他们的实验表明, 在理论计算数据与实验数据的一致性有保证的情况下, 机器学习可以同时利用实验数据和理论计算数据, 更加高效地进行模拟计算. Ward实验组^[64]用决策树模型去训练和学习无机晶体化合物的结构对形成焓的影响, 从而预测在无法得知完全平衡的晶体结构的情况下的材料形成能. 通过对比可以发现, 完全使用机器学习方法的计算代价要远低于DFT计算.

机器学习不仅可以由DFT方法的计算结果用于训练预测性质, 还可以用于优化DFT方法本身的计算过程. 由于密度泛函理论中, 能量泛函中的动能部分, 尤其是有相互作用的电子体系的动能项一无所知, 所以需要求解Kohn-Sham方程, 而Kohn-Sham方程将有相互作用电子系统的全部复杂性都归入了交换关联相互作用泛函中, 交换关联泛函的近似处理决定了DFT计算的准确度, 而这一部分的近似计算也是DFT计算中耗时最久的一部分. 现有的近似计算方法都具有一定的局限性, 例如在高温、高压以及强磁场的环境下难以拟合出正确合理的结果. 将机器学习的方法与密度泛函理论相结合, 用机器学习的方法去改进现有的基于DFT的第一性原理计算方法, 利用机器学习的灵活性与高效性, 可以在一定程度上降低计算量, 加快计算速度. 运用机器学习的方法去寻找合适的密度泛函^[65-67], 可以快速有效地寻找到高度准确的结果, 这种采用机器学习的方法来代替现在的近似计算的方法, 对基础物理学的要求很低, 只需要提供大量的训练数据就能达到化学精度.

类似的, 目前已有的各种理论计算方法中, 各种半经验参数都可以应用机器学习的方法去训练归纳, 利用机器学习的灵活性来扩展现有方法的使用范围, 除了密度泛函理论中的密度泛函, 分子动力学中原子间势^[68-70]、分子力场^[71,72]等等都可以用机器学习的方法去拟合.

此外, 由于机器学习方法可以直接学习与建立输入与输出数据之间的映射关系, 所以我们可以推测出, 应用机器学习的方法, 理论上存在直接建立起原子结构信息与能量之间的映射关系的可能性. 运用人工神经网络或者其他算法, 把包括原子种类, 原子坐标等原子结构信息作为输入, 分子能量作为输出, 直接学习其中的映射关系. 在现有的理论知识框架下, 很难直接找到具有物理意义的映射关系, 因为其复杂程度过高, 但是在机器学习的方法下, 这些复杂关系可以当作“黑箱”, 通过调整机器学习的模型参数, 跳过这些步骤, 直接训练学习输入与输出之间的映射关系. 这种方法对于算法模型中的特征选取以及参数设定等方面具有较高的难度, 具体方法还有待进一步研究.

4.3 与高通量筛选方法相结合

机器学习与高通量筛选也有一些互通性, 将高通

量筛选与机器学习结合起来也许会进一步提升理论模拟计算的效率. 高通量筛选^[73]的方法可以在短时间内同时对上千种可能的材料进行过滤, 寻找其中符合特定性能标准的体系供实验参考. 目前, 高通量筛选的计算步骤主要由以下几个部分构成: 从各种数据库中选择合适的数据; 生成输入数据; 选择合适的算法进行性能计算模拟; 产生输出文件并进行数据分析; 将处理过的数据添加到现有数据库中进行扩充, 如图8所示. 高通量筛选方法筛选出来的材料依然很难达到直接投入到实际应用的水平, 究其原因, 主要是因为筛选条件和筛选方法还有待完善和提高. 筛选条件的选择是影响高通量筛选的决定性因素, 对于新材料的筛选, 不仅需要满足各种热力学和动力学上的理论性质, 同时还要考虑实际应用过程偏离理论的影响, 包括材料中可能存在的各种缺陷, 材料的尺寸效应、表面效应以及掺杂等, 这对于筛选条件的选择, 提出了极高的要求. 筛选方法方面, 主要包括两种方法: 一种是基于原子间势的方法, 包括分子动力学方法, 而另一种是电子结构方法, 尤其是基于密度泛函理论的计算方法. 这些方法存在的不足以及机器学习提供的改进思路在前一节中已经有所提及.

高通量筛选方法和机器学习在一定程度上有着一定的相似性. 高通量筛选在新材料的设计过程中表现出

图8 (网络版彩色)高通量筛选的流程图

Figure 8 (Color online) Flow chart of high-throughput screening.

来的主要特点为自动化程度高, 可同时处理的样本量大, 另外还具有可并行、可扩展的优点, 而机器学习最为人工智能领域的一个分支, 其自动化程度更是有过之而无不及, 此外, 机器学习的算法可处理的数据量也很大, 甚至, 机器学习对于数据量还存在一定的依赖性, 数据量过小反而可能导致机器学习出现过拟合的现象而使得方法的有效性降低^[61], 同时, 机器学习方法里也包含增量学习^[47,74]和在线学习^[75-77]的方法, 所以机器学习同样可以具有可并行、可扩展的特点. 二者具有相似的特点, 所以这两种方法具有结合起来同时应用到新材料的开发之中的潜质, 值得去尝试.

另外, 针对高通量筛选与机器学习两者各自的缺陷, 将这两种方法结合起来应用可以起到互利互补, 扬长避短的效果. 高通量筛选中的筛选方法可以不局限于基于密度泛函理论的第一性原理计算方法, 也可以用数据驱动的机器学习进行结合与补充, 以达到进一步加快和优化算法的目的; 高通量筛选的筛选条件的制定需要研究者对于材料的结构与性质之间的关联有更深刻的理解和认知, 这一步可以由机器学习的方法去探索结构性质的关联性, 为研究者提供指导意见. 目前的高通量筛选方法已经形成了巨大的数据库, 这些数据正好可以用作机器学习方法中用于提升算法的训练数据. 高通量筛选与机器学习的相似性与互补性为二者的结合奠定了坚实的基础, 如果将这两种方法程序性的结合起来, 有希望让新材料的筛选工作得到进一步的加速与提升.

在材料科学领域, 高通量筛选主要还是依靠基于DFT的第一性原理计算, 先前已经有一些工作指出了结合这两种方法的可能性^[78]. 例如, Tanaka实验组提出了基于密度泛函理论构建材料大数据, 然后将其用作机器学习的训练集^[62], 该方法就体现了将高通量筛选方法与机器学习相结合的思想. 有实验组通过高通量筛选与机器学习的方法相结合应用于新型有机发光二极管分子的筛选^[79], 通过机器学习方法和DFT计算的结合, 从160万种候选结构中筛选出数千个具有研究价值的材料体系. 还有一些实验组利用机器学习的高效性, 将其与高通量筛选相结合, 应用于快速准确地识别可以用于CO₂捕获的金属有机框架材料^[80]. 值得令人鼓舞的是机器学习的应用使得计算时间降低了一个数量级, 能实现大型结构库的快速搜索. 随着计算机计算能力的发展以及人类对于机器学习方法的深入

研究, 机器学习必将在材料科学中大显身手.

5 总结与展望

机器学习是人工智能领域一个新兴的分支, 在材料科学领域具有很大的潜力, 由于在实际情况下处理复杂体系的问题中的灵活性、准确性与泛化能力, 机器学习值得研究者的充分重视. 机器学习的应用具有很强的跨学科性, 机器学习在材料科学中的应用对象十分广泛, 包括金属材料、无机氧化物材料、有机材料和各类功能材料. 各种各样的相关研究表明, 机器学习可以作为一种快速、准确的工具用于材料科学领域. 随着理论和方法的进一步发展, 机器学习必将在材料科学领域中拥有更加深入、广泛的应用.

然而, 机器学习的方法仍然有一些局限性, 主要是因为机器学习是一种数据驱动的方法, 机器学习对于数据的依赖性比较强, 在数据比较有限、数据集比较小的情况下, 机器学习的方法容易出现过拟合的现象, 大大降低方法的泛化能力. 另外, 机器学习做出的预测或者决策主要是基于经典概率论与数理统计基础的, 其中的物理和化学意义仍然有待进一步的研究与解释. 由于目前机器学习在材料研发中的应用还处于初级阶段, 未来还有很多工作需要进一步深化^[81]. 包括: (1) 结合理论和实验建立可靠的包含结构信息、功能信息和表征信息的大数据库; (2) 发展更加快速准确的学习算法; (3) 将机器学习有机地与密度泛函理论、分子动力学、蒙特卡罗等其他物理学方法相结合, 从寻找具有高普适性的电子-电子间的交换关联泛函和原子-原子间的有效相互作用势函数, 再到超规模筛选新型材料, 有机地在空间尺度和时间尺度上实现从微观到宏观的跨越; (4) 利用机器学习加强“以材料性能为导向的逆向设计”, 包括: 高容量、高倍率特性、高稳定性的电池材料; 高ZT值的热电材料; 高转换效率的太阳能材料和CO₂还原材料; 新型的拓扑量子材料(拓扑点材料、拓扑线材料及拓扑面材料), 以及具有特殊力学性能和电磁特性的超材料(Meta-Materials). 对于这类问题的研究涉及由特定的物理化学性能反推到材料中原子的空间构型排布, 而常规的方法不是十分有效, 因此可望机器学习能有力推动该领域的发展; (5) 利用机器学习加速“超原子组装材料”的设计. 在新材料的研究中超原子组装材料一直是一个十分重要而富有挑战性的课题. 在这类材料中, 超原

子可以是原子团簇、分子团簇、纳米结构. 除了不同化学组分、不同尺寸、不同取向、不同配体所带来的复杂性外, 由于超原子内部和超原子之间的不同自由度的存在, 就使得电子-电子、电子-声子、电子-光子之间的耦合相互作用变得非常复杂, 加上体系元包中所含的原子数巨大, 这就给传统的计算模拟方法带来

了极大的困难. 这也就是自从20世纪80年代初到现在这一领域发展缓慢的原因之一, 机器学习的外观为这一领域的发展带了新的生机. 总之, 随着高新技术对新材料要求的不断提高, 以及计算机技术和计算方法本身的不断发展, 机器学习必将会在新材料研发中大显身手, 甚至可能带来颠覆性的变革.

参考文献

- 1 Silver D, Huang A, Maddison C J, et al. Mastering the game of Go with deep neural networks and tree search. *Nature*, 2016, 529: 484–489
- 2 Silver D, Schrittwieser J, Simonyan K, et al. Mastering the game of Go without human knowledge. *Nature*, 2017, 550: 354–359
- 3 Shin H C, Roth H R, Gao M, et al. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Trans Med Imag*, 2016, 35: 1285–1298
- 4 Cambria E, White B. Jumping NLP curves: A review of natural language processing research. *IEEE Comput Intell Mag*, 2014, 9: 48–57
- 5 Liu H, Xu C, Liang J. Dependency distance: A new perspective on syntactic patterns in natural languages. *Phys Life Rev*, 2017, 21: 171–193
- 6 Tsai C W, Lai C F, Chiang M C, et al. Data mining for internet of things: A survey. *IEEE Commun Surv Tutor*, 2014, 16: 77–97
- 7 Argall B D, Chernova S, Veloso M, et al. A survey of robot learning from demonstration. *Robotics Autonomous Syst*, 2009, 57: 469–483
- 8 Cully A, Clune J, Tarapore D, et al. Robots that can adapt like animals. *Nature*, 2015, 521: 503–507, arXiv: 1407.3501
- 9 Kononenko I. Machine learning for medical diagnosis: History, state of the art and perspective. *Artif Intell Med*, 2001, 23: 89–109
- 10 Kononenko I. Inductive and bayesian learning in medical diagnosis. *Appl Artif Intell*, 1993, 7: 317–337
- 11 Feng N, Wang H J, Li M. A security risk analysis model for information systems: Causal relationships of risk factors and vulnerability propagation analysis. *Inf Sci*, 2014, 256: 57–73
- 12 Lu W, Xiao R, Yang J, et al. Data mining-aided materials discovery and optimization. *J Materiomics*, 2017, 3: 191–201
- 13 Ramprasad R, Batra R, Pilania G, et al. Machine learning in materials informatics: Recent applications and prospects. *npj Comput Mater*, 2017, 3: 54, arXiv: 1707.07294
- 14 de Mantaras R L, Armengol E. Machine learning from examples: Inductive and Lazy methods. *Data Knowl Eng*, 1998, 25: 99–123
- 15 Jordan M I, Mitchell T M. Machine learning: Trends, perspectives, and prospects. *Science*, 2015, 349: 255–260
- 16 Mitchell T. Machine Learning. New York: McGraw-Hill Education, 1997
- 17 Wu X, Kumar V, Ross Quinlan J, et al. Top 10 algorithms in data mining. *Knowl Inf Syst*, 2008, 14: 1–37
- 18 Quinlan J R. Induction of Decision Trees. In: Machine Learning. Dordrecht: Kluwer Academic Publishers-Plenum Publishers, 1986
- 19 Safavian Sr L D. A survey of decision tree classifier methodology. *IEEE Trans Syst Man And Cybernet*, 1991, 21: 660–674
- 20 Zhang Z. Naïve Bayes classification in R. *Ann Transl Med*, 2016, 4: 241
- 21 Burges C J. A Tutorial on Support Vector Machines for Pattern Recognition. In: Data Mining And Knowledge Discovery. Dordrecht: Kluwer Academic Publishers, 1998
- 22 Wloka J. Kernel functions and nuclear spaces. *Bull Am Math Soc*, 1965, 71: 720–723
- 23 Han J W. Data Mining: Concepts and Techniques. Beijing: China Machine Press, 2007 [韩家炜. 数据挖掘概念与技术. 北京: 机械工业出版社, 2007]
- 24 Han J W, Pei J, Yin Y W, et al. Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach. In: Data Mining And Knowledge Discovery. Dordrecht: Kluwer Academic Publishers, 2004
- 25 Goh G B, Hodas N O, Vishnu A. Deep learning for computational chemistry. *J Comput Chem*, 2017, 38: 1291–1307
- 26 Zhang G P. Neural networks for classification: A survey. *IEEE Trans Syst Man Cybernet C*, 2000, 30: 451–462
- 27 LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*, 2015, 521: 436–444
- 28 Liu Y, Zhao T, Ju W, et al. Materials discovery and design using machine learning. *J Materiom*, 2017, 3: 159–177
- 29 Raccuglia P, Elbert K C, Adler P D F, et al. Machine-learning-assisted materials discovery using failed experiments. *Nature*, 2016, 533: 73–76
- 30 Ferguson A L. Machine learning and data science in soft materials engineering. *J Phys-Condens Matter*, 2018, 30: 043002

- 31 Eremin R A, Zolotarev P N, Ivanshina O Y, et al. Li(Ni,Co,Al)O₂ cathode delithiation: A combination of topological analysis, density functional theory, neutron diffraction, and machine learning techniques. *J Phys Chem C*, 2017, 121: 28293–28305
- 32 Wang X, Xiao R, Li H, et al. Quantitative structure-property relationship study of cathode volume changes in lithium ion batteries using *ab-initio* and partial least squares analysis. *J Materiom*, 2017, 3: 178–183
- 33 Gaultois M W, Oliynyk A O, Mar A, et al. Perspective: Web-based machine learning models for real-time screening of thermoelectric materials properties. *APL Mater*, 2016, 4: 053213, arXiv: 1502.07635
- 34 Seko A, Togo A, Hayashi H, et al. Prediction of low-thermal-conductivity compounds with first-principles anharmonic lattice-dynamics calculations and bayesian optimization. *Phys Rev Lett*, 2015, 115: 205901, arXiv: 1506.06439
- 35 Li H, Zhang Z, Liu Z. Application of artificial neural networks for catalysis: A review. *Catalysts*, 2017, 7: 306
- 36 Ulissi Z W, Tang M T, Xiao J, et al. Machine-learning methods enable exhaustive searches for active bimetallic facets and reveal active site motifs for CO₂ reduction. *ACS Catal*, 2017, 7: 6600–6608
- 37 Fernandez M, Bilić A, Barnard A S. Machine learning and genetic algorithm prediction of energy differences between electronic calculations of graphene nanoflakes. *Nanotechnology*, 2017, 28: 38LT03
- 38 Lopez E, Gonzalez D, Aguado J V, et al. A manifold learning approach for integrated computational materials engineering. *Arch Computat Methods Eng*, 2018, 25: 59–68
- 39 Wang Y, Zhang W, Chen L, et al. Quantitative description on structure-property relationships of Li-ion battery materials for high-throughput computations. *Sci Tech Adv Mater*, 2017, 18: 134–146
- 40 Attarian Shandiz M, Gauvin R. Application of machine learning methods for the prediction of crystal system of cathode materials in lithium-ion batteries. *Comput Mater Sci*, 2016, 117: 270–278
- 41 Fujimura K, Seko A, Koyama Y, et al. Accelerated materials design of lithium superionic conductors based on first-principles calculations and machine learning algorithms. *Adv Energy Mater*, 2013, 3: 980–985
- 42 Kireeva N, Pervov V S. Materials space of solid-state electrolytes: Unraveling chemical composition-structure-ionic conductivity relationships in garnet-type metal oxides using cheminformatics virtual screening approaches. *Phys Chem Chem Phys*, 2017, 19: 20904–20918
- 43 Jalem R, Aoyama T, Nakayama M, et al. Multivariate method-assisted *Ab Initio* study of olivine-type LiMXO₄ (main group M²⁺–X⁵⁺ and M³⁺–X⁴⁺) compositions as potential solid electrolytes. *Chem Mater*, 2012, 24: 1357–1364
- 44 Jalem R, Nakayama M, Kasuga T. An efficient rule-based screening approach for discovering fast lithium ion conductors using density functional theory and artificial neural networks. *J Mater Chem A*, 2014, 2: 720–734
- 45 Jalem R, Kimura M, Nakayama M, et al. Informatics-aided density functional theory study on the Li ion transport of tavorite-type LiMTO₄F (M³⁺–T⁵⁺, M²⁺–T⁶⁺). *J Chem Inf Model*, 2015, 55: 1158–1168
- 46 Li X Y, Shu X, Shen J W, et al. An on-board remaining useful life estimation algorithm for lithium-ion batteries of electric vehicles. *Energies*, 2017, 10: 691
- 47 Liu D, Zhou J, Pan D, et al. Lithium-ion battery remaining useful life estimation with an optimized Relevance Vector Machine algorithm with incremental learning. *Measurement*, 2015, 63: 143–151
- 48 Sun F, Li X, Liao H, et al. A Bayesian least-squares support vector machine method for predicting the remaining useful life of a microwave component. *Adv Mech Eng*, 2017, 9: 1–9
- 49 Song Y, Liu D, Yang C, et al. Data-driven hybrid remaining useful life estimation approach for spacecraft lithium-ion battery. *MicroElectron Reliability*, 2017, 75: 142–153
- 50 Hu X, Li S E, Yang Y. Advanced machine learning approach for lithium-ion battery state estimation in electric vehicles. *IEEE Trans Transp Electrific*, 2016, 2: 140–149
- 51 Li F, Xu J. A new prognostics method for state of health estimation of lithium-ion batteries based on a mixture of Gaussian process models and particle filter. *MicroElectron Reliability*, 2015, 55: 1035–1045
- 52 Hu C, Jain G, Schmidt C, et al. Online estimation of lithium-ion battery capacity using sparse Bayesian learning. *J Power Sources*, 2015, 289: 105–113
- 53 Zhu Y. State of health estimation based on OS-ELM for lithium-ion batteries. *Int J Electrochem Sci*, 2017, 6895–6907
- 54 Schoenholz S S, Cubuk E D, Sussman D M, et al. A structural approach to relaxation in glassy liquids. *Nat Phys*, 2016, 12: 469–471
- 55 Cubuk E D, Schoenholz S S, Rieser J M, et al. Identifying structural flow defects in disordered solids using machine-learning methods. *Phys Rev*

- [Lett](#), 2015, 114: 108001, arXiv: [1409.6820](#)
- 56 Han Y F, Zeng W D, Shu Y, et al. Prediction of the mechanical properties of forged Ti-10V-2Fe-3Al titanium alloy using FNN. [Comput Mater Sci](#), 2011, 50: 1009–1015
- 57 Zhang W, Sun P, Sun S. A precise theoretical method for high- throughput screening of novel organic electrode materials for Li-ion batteries. [J Materiom](#), 2017, 3: 184–190
- 58 Jain A, Ong S P, Hautier G, et al. Commentary: The materials project: A materials genome approach to accelerating materials innovation. [APL Mater](#), 2013, 1: 011002
- 59 Philip Chen C L, Zhang C Y. Data-intensive applications, challenges, techniques and technologies: A survey on big data. [Inf Sci](#), 2014, 275: 314–347
- 60 Chen X W, Lin X T. Big data deep learning: Challenges and perspectives. [IEEE Access](#), 2014, 2: 514–525
- 61 Zhou L, Pan S, Wang J, et al. Machine learning on big data: Opportunities and challenges. [Neurocomputing](#), 2017, 237: 350–361
- 62 Takahashi K, Tanaka Y. Material synthesis and design from first principle calculations and machine learning. [Comput Mater Sci](#), 2016, 112: 364–367
- 63 Hautier G, Fischer C C, Jain A, et al. Finding nature’s missing ternary oxide compounds using machine learning and density functional theory. [Chem Mater](#), 2010, 22: 3762–3767
- 64 Ward L, Liu R, Krishna A, et al. Including crystal structure attributes in machine learning models of formation energies via Voronoi tessellations. [Phys Rev B](#), 2017, 96: 024104
- 65 Brockherde F, Vogt L, Li L, et al. Bypassing the Kohn-Sham equations with machine learning. [Nat Commun](#), 2017, 8: 872, arXiv: [1609.02815](#)
- 66 Snyder J C, Rupp M, Hansen K, et al. Finding density functionals with machine learning. [Phys Rev Lett](#), 2012, 108: 253002, arXiv: [1112.5441](#)
- 67 Hegde G, Bowen R C. Machine-learned approximations to density functional theory hamiltonians. [Sci Rep](#), 2017, 7: 42669, arXiv: [1609.05737](#)
- 68 Deringer V L, Csányi G. Machine learning based interatomic potential for amorphous carbon. [Phys Rev B](#), 2017, 95: 094203, arXiv: [1611.03277](#)
- 69 Podryabinkin E V, Shapeev A V. Active learning of linearly parametrized interatomic potentials. [Comput Mater Sci](#), 2017, 140: 171–180
- 70 Li Z, Kermode J R, De Vita A. Molecular dynamics with on-the-fly machine learning of quantum-mechanical forces. [Phys Rev Lett](#), 2015, 114: 096405
- 71 Chmiela S, Tkatchenko A, Sauceda H E, et al. Machine learning of accurate energy-conserving molecular force fields. [Sci Adv](#), 2017, 3: e1603015, arXiv: [1611.04678](#)
- 72 Betz R M, Walker R C. Paramfit: Automated optimization of force field parameters for molecular dynamics simulations. [J Comput Chem](#), 2015, 36: 79–87
- 73 Green M L, Takeuchi I, Hatrick-Simpers J R. Applications of high throughput (combinatorial) methodologies to electronic, magnetic, optical, and energy-related materials. [J Appl Phys](#), 2013, 113: 231101
- 74 Wang Y R, Yang L, Chen M H, et al. Incremental learning algorithm of least square twin KSVC. In: Proceedings of 2016 8th International Conference on Computer and Automation Engineering. EDP Sciences, 2016
- 75 Souza R C S N P, Leite S C, Borges C C H, et al. Online algorithm based on support vectors for orthogonal regression. [Pattern Recogn Lett](#), 2013, 34: 1394–1404
- 76 Saad D, Solla S A. On-line learning in soft committee machines. [Phys Rev E](#), 1995, 52: 4225–4243
- 77 Wang X, Han M. Online sequential extreme learning machine with kernels for nonstationary time series prediction. [Neurocomputing](#), 2014, 145: 90–97
- 78 Saal J E, Kirklin S, Aykol M, et al. Materials design and discovery with high-throughput density functional theory: The open quantum materials database (OQMD). [JOM](#), 2013, 65: 1501–1509
- 79 Gómez-Bombarelli R, Aguilera-Iparraguirre J, Hirzel T D, et al. Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. [Nat Mater](#), 2016, 15: 1120–1127
- 80 Fernandez M, Boyd P G, Daff T D, et al. Rapid and accurate machine learning recognition of high performing metal organic frameworks for CO₂ capture. [J Phys Chem Lett](#), 2014, 5: 3056–3060
- 81 De Luna P, Wei J, Bengio Y, et al. Use machine learning to find energy materials. [Nature](#), 2017, 552: 23–27

Applying machine learning to accelerate new materials development

WU Wei^{1,2} & SUN Qiang^{1,2*}

¹*Department of Materials Science and Engineering, College of Engineering, Peking University, Beijing 100871, China;*

²*Center for Applied Physics and Technology, Peking University, Beijing 100871, China*

Materials are not only the foundation of the national economy, but also the carrier of high technology. It has become a research hotspot in the world to overcome the conventional methods and apply new methods to accelerate the development of new materials. Propelled by the great success in other fields, data-driven informatics methods begin to emerge as a new technique in material science. Machine learning, as a representative of data-driven methods, has received extensive attention in various fields. Machine learning is an interdisciplinary science that combines computer science, statistics, computational mathematics and engineering. In the field of materials science, machine learning methods show faster calculation speed and higher prediction accuracy compared with conventional theoretical computational simulations based on solving physical or chemical fundamental equations. Machine learning is an effective addition to the existing theoretical calculation methods and significantly increases the efficiency of materials computational simulation work. Furthermore, it also works for some systems or problems that the traditional theoretical calculation methods fail to solve. This approach could also enable targeted material design and development. This review would provide a brief overview on the fundamentals of machine learning, several typical algorithms in machine learning and the applications in materials science, and discuss the future challenges in this field.

new materials, machine learning, computation and simulation

PACS: 02.70.-c, 61.50.Ah, 71.15.-m

doi: [10.1360/SSPMA2018-00073](https://doi.org/10.1360/SSPMA2018-00073)