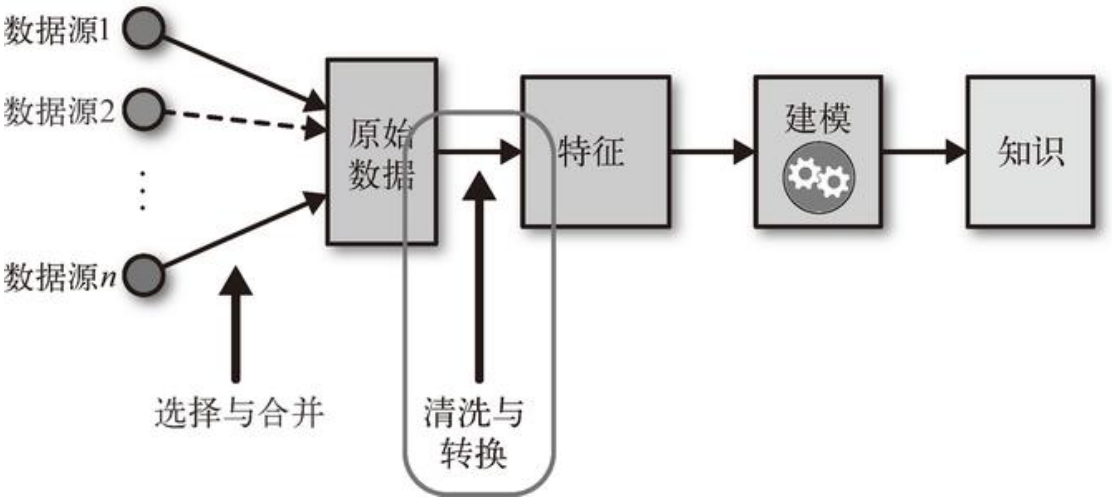
浅谈什么是材料数据机器学习的特征

11 月份 MatCloud+顺利开展了【催化领域的材料计算和机器学习公开课】，在第三节课时，对于机器学习在催化领域的研究大家很感兴趣，有同学提问：什么是材料数据的特征？应该如何选择等问题，今天小编就针对以上问题做了些简单的整理分享给大家，快来一起看看吧。

机器学习的流程包含以下几部分：数据收集、数据预处理、特征值工程、训练模型、性质预测。



其中，**特征工程**是基于领域知识设计感兴趣系统的数字指纹的过程。

构建机器学习模型时，识别全面和合适的特征是最重要和最具挑战性的步骤，因为它与模型的准确性有非常重要的关系。因此，将机器学习应用于的催化研究中时，从数百种材料特性中选择催化反应描述符作为模型特征至关重要。

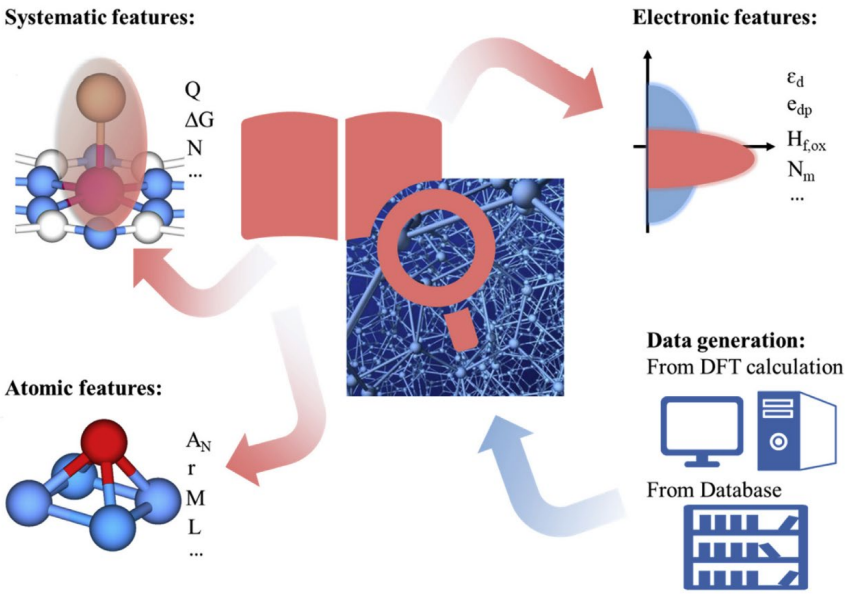
一、 特征值的选择标准

一般，可以采用几种方法来选择合适的特征作为模型的输入数据：

- ◆ 特征应该可以单独代表化学系统的原子和电子结构的一个方面；
- ◆ 特征需要包括一些能够有效描述活性位点的局部化学环境的描述符；
- ◆ 特征应该通过很少的 DFT 计算或可以直接从可用的数据库查询中获得，以提高机器学习方法的效率；
- ◆ 它们应该在物理上应该是直观的，以保证模型的准确性。

二、 催化领域如何选择特征值

基于以上方法，来自 DFT 计算或材料数据库的特征或描述符可以分为以下三类：



- ✧ **原子特征**：例如原子序数 (AN)、原子半径 (r)、相对原子质量 (M) 和键长 (L)；
- ✧ **电子特征**：例如 d 轨道和 p 轨道的电子数 (edp)、d 带中心 (εd)、氧化物形成焓 (Hf,ox)、泡利电负性 (Nm)、电子亲和势 (χ) 和第一电离能 (Im)原子；
- ✧ **系统特征**：例如系统中某一部分的电荷转移 (Q)、吸收能 (ΔG) 和有效配位数 (N)。
- ✧ 单个原子或化学系统的单个部分的这些单独特征可以组合成许多其他特征。

三、 特征值是否有意义

以上是关于将机器学习用于催化研究领域时选择特征值的一些建议，除此之外我们还需

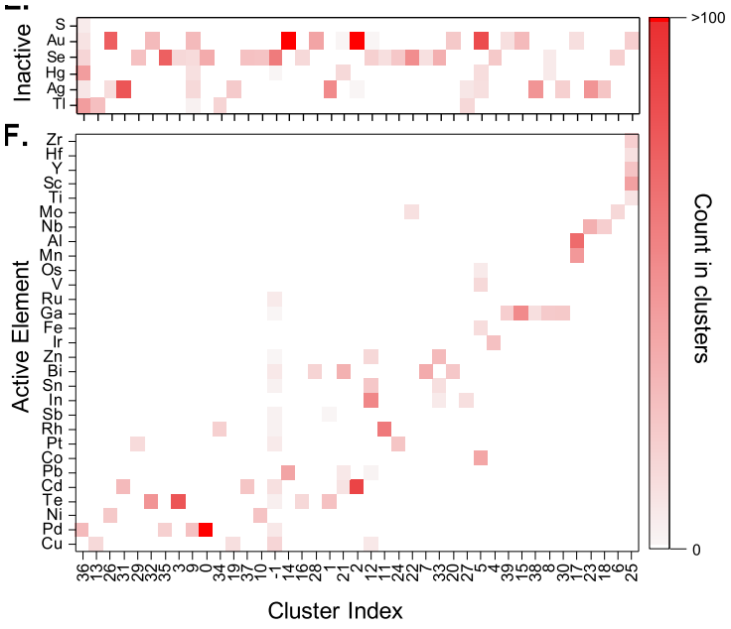
要选择有意义的特征输入机器学习的算法和模型进行训练。

通常来说，从两个方面考虑来选择特征：

- ✓ 特征是否发散：不发散的特征其方差更接近于 0，即说样本在此特征上几乎无差异，因而这个特征对于样本的区分也无效。
- ✓ 特征与目标的相关性：与目标相关性高的特征，应当优选选择。

四、 特征值的数量

除了特征值的选择，特征的数量也很重要，因为特征太少的模型通常会过度简化变量之间的关系(拟合不足)，而特征太多的模型在预测时通常表现出很大的高可变性(过度拟合)。并且值得注意的是，在材料数据集上训练的机器学习模型通常适用于类似系统，但在其他材料系统中表现较差，这提醒我们在特征工程中要多加考虑。这里就涉及到了材料数据特征提取和特征筛选的问题，我们今后会给大家陆续介绍。



关系热力图

今天我们为简单地介绍了下什么是材料数据机器学习的特征，**后期还会发布更多关于计算模拟和机器学习的各种干货，如果您想了解更多请持续关注我们。**

更多 Matcloud+教程可关注 **b 站迈高科技**。
更多动态请关注**迈高科技微信公众号**