# 计算材料学论文读后有感

今日阅读过一篇《基于DFT辅助机器学习的铁基双金属硫属化合物磁性发现》的外文文献，大受启发。正好本人在机器学习上略有了解，随即手作文写下从中学到的知识。

## 研究背景

而今永磁体随着能源领域发展需求越来越大，传统的永磁体原材料稀土元素处于耗尽的临界状态，这使得人们急需一种新的磁性材料以不被稀土元素的稀缺制约。研究人员也进行过不少实验，取得了不小的突破。

其中铁基硫族化合物脱颖而出，根据近日来针对不同硫元的铁磁性、铁磁性和反铁磁性的报告，发现通过改变金属元素和硫族元素来改变过渡金属硫族元素的元素组成，可能会揭示一种新的磁性材料形式。

然而，研究所有可能的硫属化合物组成是一个公开的挑战。实验研究涉及到这些材料的合成和表征，这是非常昂贵和耗时的。基于第一原理密度泛函理论(DFT)的预测计算，明确包括有效单粒子图像中的电子-电子相互作用，考虑到硫族化合物的多种组成构型，在数值上也具有挑战性。在这样的情况下，一个完善的数据驱动方法可以提供一个更快和计算成本效益的替代方案，能够使研究中的人工筛选成本大大减少。近年来，也踊跃了不少使用DFT和实验数据训练的机器学习模型的报道，突出了DFT在数据驱动的材料科学领域的重要性。

## 研究思路

预测铁基双金属硫属化合物FexAyB的组成依赖磁性。A代表Ni、Co、Cr、Mn, B代表S、Se、Te, x、y分别代表各自元素的浓度。先使用密度泛函理论(DFT)计算生成了一个包含4348种铁基双金属硫属化合物FexAyB组成构型的结构数据集。该数据集随后被用于训练各种ML算法，如线性回归、支持向量回归、随机森林、决策树、k近邻、极端梯度增强和人工神经网络。基于十倍交叉验证，我们选择了六个最好的机器学习算法来开发一个基于堆叠泛化的集成模型，用于预测双金属硫属化合物中的磁性。

## 研究方法

首先，他们生成了一个包含4348个结构的数据集，代表各种组成构型，然后进行DFT21计算并获得了单元细胞的磁化强度。然后，他们进行了特征工程，其中他们采用了一组适合于描述硫属化合物磁化的描述符(特征)。他们将数据集分为训练集和测试集。随后，他们使用交叉验证和网格搜索方法训练模型，以确定表现最佳的模型。最后，他们在(独立的)测试集上测试了我们提出的模型的性能。

## 创建数据集

首先采用了DFT计算的第一原则，构建了一个由两个Fe原子和两个S原子组成的六方(空间群p63/mmc)铁-硫化物(FeS)的原始细胞，利用优化后的晶格参数，我们将六方FeS的原始单元扩展为更大的单元，由16个Fe和16个s原子组成。为了描述结构的局部几何结构，通过S1、S2、S3和S4指定了4个原子位点。这些位置S1, S2, S3和S4被分配了一个数字，这取决于在这些原子位置上有多少铁原子被取代，同时保持硫离子浓度不变。根据取代的数量和位置，生成了4348个不同组成的双金属硫族化合物结构，随后，进行自旋极化DFT计算，计算每种成分的磁化强度，这些磁化强度将作为预测值归入机器学习的训练集。而特征值选择了12个，其中8个特征值描述了金属元素的浓度，如金属Fe, Ni, Co, Cr, Mn，以及硫化物元素S, Se和Te。剩下4个特征值描述了替换位点S1、S2、S3和S4的位置。这些描述符的选择是由于取代硫族化合物的磁顺序也依赖于取代位点。（根据s1234的取代情况可以推出硫族化合物的取代位点，从而可以做到减少一个特征值，即给数据降一维。）最后将预测值与特征值合并为所需的数据集，而后85:15的比例随机分为训练集和测试集。

## 模型选择和算法训练

训练了七种不同的机器学习算法：线性回归(LR)，支持向量回归(SVR)，随机森林(RF)，决策树(DT),K近邻算法(KNN)，极端梯度增强(XGBoost)和人工神经网络(ANN)。为了找到模型的最优超参数，我们对训练集进行了广泛的十倍交叉验证，并对不同的超参数组合进行了网格搜索

#### 预测评估指标

为了评估不同机器学习回归模型的性能，使用了三个评估指标:均方误差(MSE)、平均绝对误差(MAE)和决定系数(R2)。R2是回归模型中的一种统计度量，它表示从自变量中可预测的因变量方差的比例，它本身并不能衡量预测的准确性，我们将这个指标与MSE和MAE结合使用来衡量我们研究中使用的回归模型的性能。

#### 结果和结论

首先，我们对DFT数据集进行了一些探索性数据分析。结果表明，与含有Te的硫族化合物相比，含有S和Se的双金属硫族化合物具有更高的磁矩，揭示了FeS和FeSe表现出比FeTe更强的磁化能力。此外，我们发现含有Cr和S的Fe -硫族化合物比含有其他过渡金属元素(Ni、Co或Mn)和硫族元素(Se和Te)的Fe -硫族化合物具有更高的磁矩。当含有S或Se的硫属化合物中Cr或Mn浓度增加时，磁矩也明显增加。此外，硫族化合物中过渡金属元素的取代位也会影响目标值(磁矩)。

## 基础学习器和元学习器的十倍交叉验证结果

通过比较平均MSE、平均MAE和平均R2，发现RF（随机森林）模型表现良好。除了LR方法外，所有模型都表现得相当好，这说明预测目标和特征关系更偏向为非线性关系。用嵌套交叉验证，使用各个基本模型的输出来训练LR、RF和XGB，并重新计算平均MSE、平均MAE和平均R2。测试集验证结果：MSE、MAE和R2分别为1.655 (μ B) (2)、0.546 μB和0.922，与10倍交叉验证值1.29 (μ B)2、0.50 μ B和0.94相比，接近。有着较好的泛化性。

## 发现参阅论文中的错误及解决方法

在运用StandardScaler()类将训练集进行预处理的步骤中，他们把所有训练集X\_train[cols\_to\_scale]使用scaler.transform方法进行了转换，而后将它们运用在交叉验证（cross\_val\_score）或是带交叉验证的网格搜索（GridSearchCV）中。但是，对于交叉验证步骤来说，它们的测试集（也称为开发集）也被包含在上面所被转换的训练集中。如此，这些开发集实际上在数据预处理时就发生了信息泄露！这会使得在进行交叉验证过程中得到过于乐观的结果，从而可能选择次优的模型参数。所以，我们应该寻找一种方法，使得任何从数据集中提取信息的处理过程都应该仅用于数据集的训练部分。

若要使任何从数据集中提取信息的处理过程都应该仅用于数据集的训练部分，最方便的是使用sklearn中的管道函数（from sklearn.pipeline import make\_pipeline）将多个估计器一起封装起来，而后嵌套在cross\_val\_score或GridSearchCV函数中，这样在交叉验证步骤中，训练集会先被分成训练部分和验证部分，而后仅有训练部分进行pipeline中的第一个估计器进行预处理，然后再传递到下一个评估器中，这样每一个交叉验证步骤中验证集就不会因为预处理过程而被提早泄露出去！

运行优化后发现，其他类型的模型经测试后发现影响不大，但在使用管道后的随机森林的网格搜索中，我重新运行了旧代码，发现旧代码中存在笔误，随机森林应该设置random\_state参数的值，以方便后来者复用时由于随机性不同而产生不同的评估结果。比如这次，源代码写的随机森林的最佳参数n\_estimators为500，而在我的电脑上运行旧代码后发现实际上为50。

在使用管道函数后我发现最优参数变为了(min\_samples\_split=13, n\_estimators=20)，未使用管道函数的预测最佳参数也为(min\_samples\_split=13, n\_estimators=20)，结合《Introducton to Machine Learning with Python》书中所说的“在交叉验证中，信息泄露的影响大小取决于预处理步骤的性质。使用测试部分来估计数据的范围通常不会产生可怕的影响，但在特征提取和特征选择中使用测试部分，则会导致结果的显著差异。”来看，这篇文章中的模拟过程的预处理性质对造成的信息泄露影响并不大，但我仍希望能在交叉验证中有使用管道函数的习惯。

综上所述，仅通过交叉验证（cross\_val\_score），使用管道函数影响不大，但是使用带有交叉验证的网格搜索（GridSearchCV）时，使用管道可能会选出更优的参数。

#### 总结

寻找比稀土元素磁体更便宜的新型磁性材料是近年来的重要课题，因为它们的应用范围从数据存储到汽车，从生物医学领域到绿色能源领域。对可能的替代磁性材料进行实验和计算研究既昂贵又耗时。在这项工作中，我们使用了一个数据驱动的框架，这将加速发现新的磁性材料。我们利用优化的FeS结构和取代技术设计了新的不同组成的双金属硫属化合物。第一原理DFT用于生成训练和测试数据，用于训练和评估了几个机器学习模型。之后，我们开发了一个堆叠的机器学习模型，它结合了几个表现最好的基础模型作为最终的预测工具。在MSE、MAE和R2值分别为1.655 (μ B)2、0.546 μ B和0.922的情况下，该模型在独立dft测试数据上显示出较高的准确性。预测数据显示，铁基双金属硫族化合物含有硫元S，过渡金属Cr浓度较高，磁矩高于其他构型，这与DFT数据一致。这项工作为发现一种由更便宜、更丰富的元素制成的新磁性材料提供了一种策略，这种材料最终将取代由稀土金属制成的昂贵的现有磁性材料。