基于机器学习的镍基单晶高温合金材料数据分析方法研究

上海大学刘悦、施思齐——2018 年 7 月 4 日

1. 整体工作计划

年度	任务	考核指标	成果形式
2017 年 7 月-2018 年 6 月	收集与存储已有镍基单晶高温合金材料计算数据；开发主动学习的多层级交互式特征选择方法，定性定量分析各种因素对性能的影响程度	实践机器学习及数据挖掘方法，为提取“数据关联”规律作准备	提出一份提取已知单晶高温合金中的数据关联规律的初步报告
2018 年 1 月-2019 年 6 月	使用机器学习方法挖掘高温合金数据，根据不同的学习目标自适应地构造出多种学习器混合预测模型，提高对性能的预测精度	开展以机器学习及数据挖掘为重点的数据关联分析	研发相关数据分析软件
2019 年 7 月-2020 年 6 月	基于数据关联分析计算，提取规律构建高通量并发式计算数据分析与管理软件	重点为发展机器学习及数据挖掘方法提出数据关联规律	研究报告及计算软件
2020 年 1 月-2020 年 12 月	研究基于规则抽取的可解释性方法，将机器学习学到的结果转为易于理解的 if-then-else 规则，提高预测方法的可解释性	发展机器学习方法用于解析关联分析	研究报告或论文

2. 最新工作进展

我组的基本任务是针对单晶高温合金材料高通量并发式集成计算的数据中

蕴含的复杂的数据关联，引入机器学习及数据挖掘方法对其内在相关性进行分析，探索高温合金材料性能预测方法、数据关联表征方法。面向高通量计算数据，发展实现复杂数据分析与管理软件。因此，本课题组紧紧围绕我们的基本任务，在过去的一年时间里，从数据、算法、应用与平台四个方面展开了系统的研究，采集到来自王院士 49 篇文献的高温合金计算数据和来自文献 / 专利的蠕变性能数据，为机器学习做好数据准备；制定了一套数据挖掘与机器学习算法规范，初步完成机器学习 / 数据挖掘算法库的研发，并针对高温合金数据特点，研发了多层级交互性特征分析方法，定性定量分析各种因素对高温合金性能的影响程度；基于以上研究，研发了高温合金机器学习演示平台。工作总览如图 1 所示。

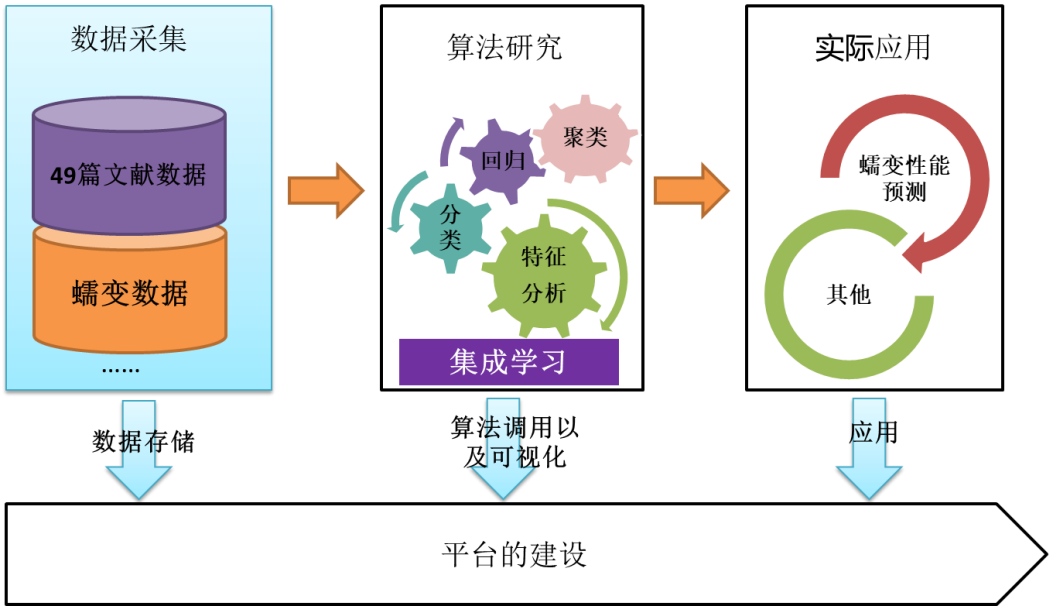


图 1 最新工作总览

2.1 高温合金数据采集和存储

我组完成了王院士 49 篇文献的精读，并对 49 篇文献进行了分类，其中研究原子占位与分配 7 篇，力学性能 12 篇，原子扩散 3 篇，错配位错、位错网络、位错运动 19 篇，结构和界面 10 篇。我们收集了每一篇文献中存在的各类数据，包括图片数据 142 条、规则数据 89 条、文献信息数据 49 条（49 篇文献）以及表结构数据 235（高温合金性能数据）；同时，也对 49 篇文献出现的高温合金性能描述因子进行了归纳和分类，主要分为以下 4 大类：合金化元素占位和相分配（晶格常数、结合能、占位形成能、原子择位能等）；错配位错、相界面裂纹、

位错运动（裂纹方向、晶格捕获上限、晶格错配度等）；扩散（空位形成能、原子迁移能）；力学性能（弹性刚度常数、杨氏模量、剪切模量等）。另外，我组也通过查阅大量文献，从文献或专利中采集了镍基单晶高温合金的蠕变性能数据453条。为了采集更多的镍基单晶高温合金性能数据，我组采纳了王院士的建议，检索和下载了每四年一次的国际高温合金会议相关论文集，并从论文中提取了大量的合金数据。同时，我组考虑到收集的所有高温合金数据来自不同的文献或专利，其中可能会存在大量不完整、不一致、异常的数据，称之为“脏数据”，严重影响到数据挖掘的效率，甚至导致数据挖掘结果的偏差，所以有必要对数据进行清洗和数据质量检查以提高数据的质量。最后，由于收集的数据类型多样，其中包括结构化数据，如二维逻辑表、半结构化数据，如文本、非结构化数据，如图片、音频等，因此我组采用了目前较流行的非关系型数据库 mongodb 和轻量级关系型数据库 mysql 对经过数据清洗和数据质量检查和分析之后的所有高温合金数据按照我组制定的数据标准规范完成了存储。

2.2 数据驱动的数据挖掘和机器学习算法研究

首先，为了所研究算法能够具有很好的平台兼容性，能更好地集成到项目平台中，我组制定了算法规范，实现了算法命名标准化、注释规范化、编码统一化，以设计出标准的 API，提高算法的可读性、可用性和可完善性；其次，按照规范实现了包括各种特征降维、回归、分类、聚类和优化算法，形成算法库；课题组还初步研究和实现了数据驱动的自动式机器学习/数据挖掘算法，我组对对现有的、基础的机器学习算法，包括无监督学习（聚类、特征降维）、有监督学习（分类和回归）进行了系统分类和逻辑层次表达，并构建了相应的算法分类树，如图2所示。该分类树综合考虑了算法本身的特性以及算法使用用户的需求，让材料领域等非计算机专家能对机器学习已有的算法有初步的认识和了解。然而，算法分类树虽然直观清晰地展示了各种机器学习算法的分类情况以及其应用场景，但并没有给出各种算法的具体实现流程，因此我们接下来提出了数据驱动的机器学习算法实现流程图，如图3所示。

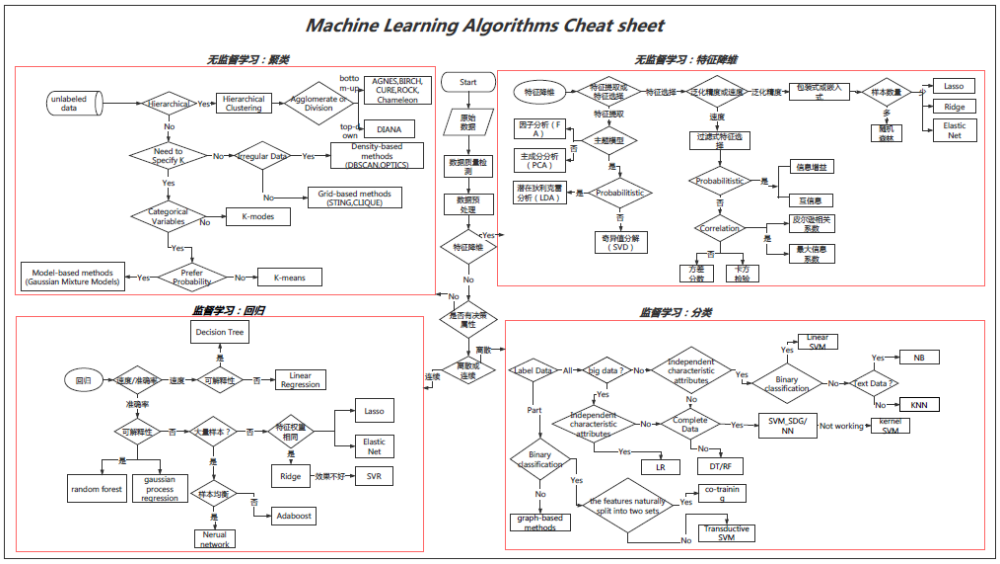


图 2 机器学习算法分类树

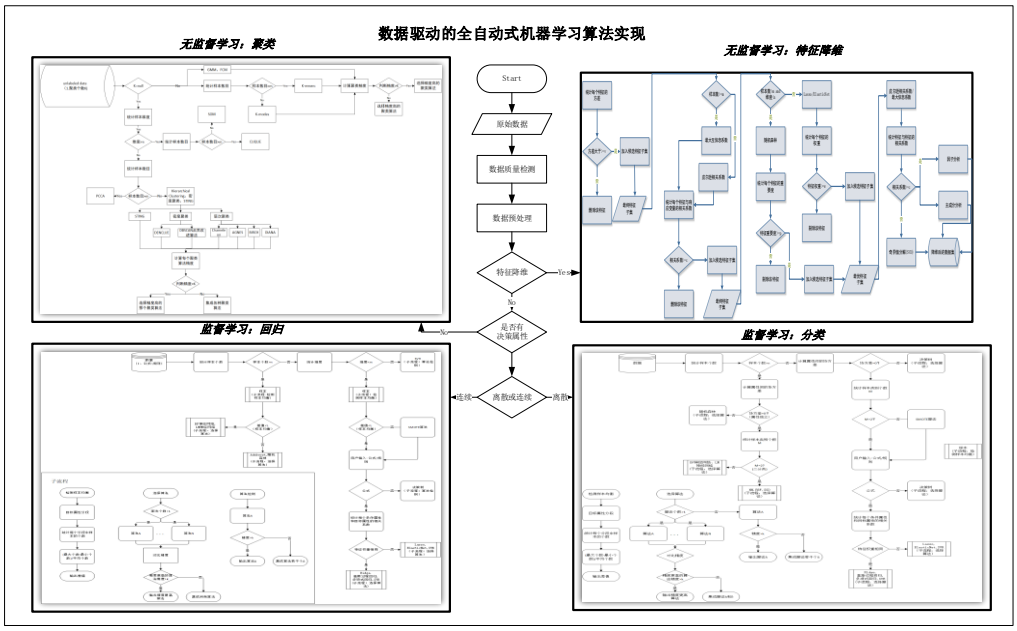


图 3 数据驱动的全自动式机器学习算法流程图·

2.3 面向高温合金数据的机器学习方法研究

2.3.1 性能预测方法研究

1) 基于聚类的最优回归集成方法

材料数据的材料数据通常呈现**小样本、高维度和分布不均匀的特点**，综合样本物理背景的复杂性，所以难以找到一种单一的模型对高维且具有一定物理复杂

性背景的小样本进行建模。针对这个问题，我们提出了基于聚类的最优回归集成学习方法。

方法框架如图 4 所示，对于输入的学习样本，首先把它投射到特征空间并利用聚类的方法把样本划分为不同的簇，其中簇内样本的特性相似、簇间样本的特性不同；针对不同的簇，依据预测的精度选择最优的回归模型，最后通过异构集成的方法集成模型。我组以蠕变预测为例，以验证该方法的有效性。

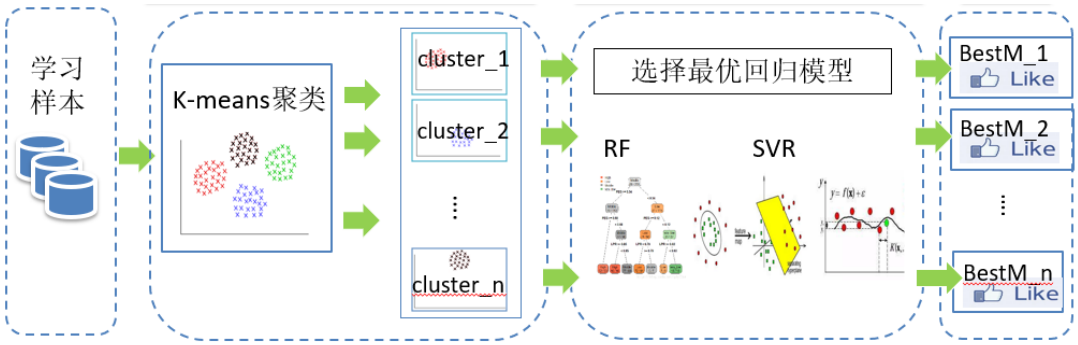


图 4 基于聚类的最优回归集成学习方法

2) 在蠕变数据中的应用

我组通过收集并整合了文献和专利的图片数据、表格数据和文本中的数据，初步得到了 453 条蠕变数据，考虑到专利数据的完整性，最后用于蠕变寿命预测的数据全部选择专利，横向看，一共 402 条蠕变数据，纵向看，它包含了 25 个维度，其中有关于成分的 14 维：Ni, Re, Co, Al, Ti, W, Mo, Cr, Ta, C, B, Y, Nb, Hf，热处理的 6 个维度：固溶处理的时间和温度，两个阶段时效处理的时间和温度，外部条件的 2 个维度：温度和应力，还有目标属性：蠕变断裂寿命、蠕变断裂应变和到达指定应变的时间。

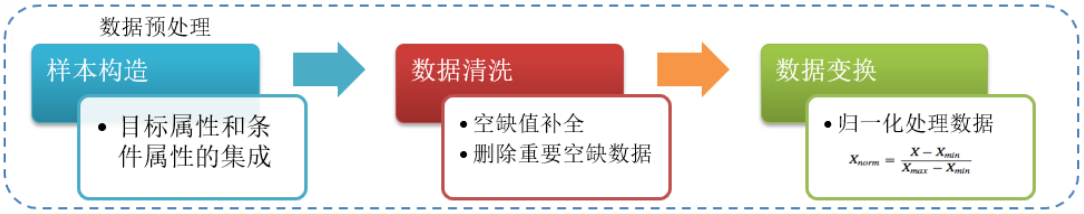


图 5 数据预处理的步骤

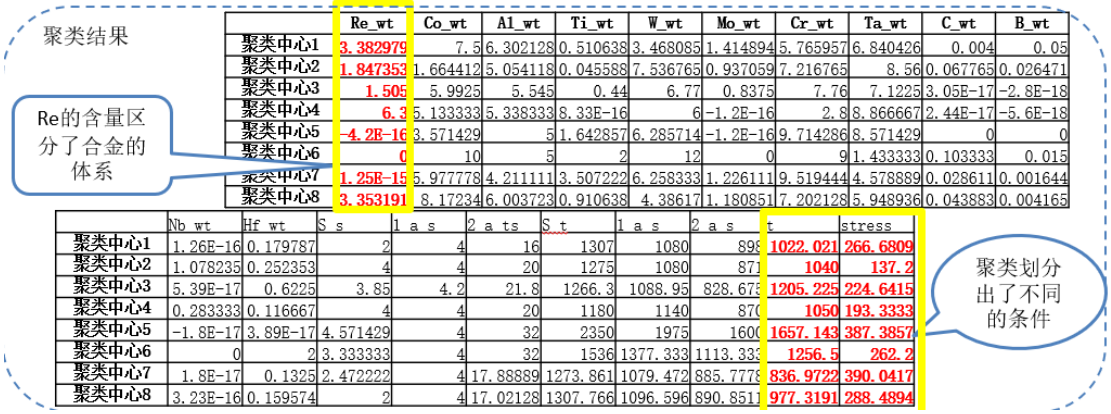
由于原始的样本存在空缺值、不一致和不同维度数值差异大等问题，不能构成学习样本，为此，我组对原始数据进行了数据预处理，如图 5 所示。首先时目标属性和条件属性的集成，目标属性选定为蠕变断裂寿命，条件属性则保留了所

有的成分、热处理和外部条件；其次对数据进行了清洗，包括了一些成分空缺值的补零处理、缺失热处理样本的删除；最后对所有数据进行了归一化处理，最终形成了可学习样本 276 条。

之后对上述的样本进行了数据质量检查和分析，通过求解各个属性的均值、方差、最小值、1/4 位数、中位数、3/4 位数、最大值、偏度和范围。通过分析，可以得出了两个重要的结论：第一，外部条件变化幅度大，包含了高温高应力、高温低应力、低温高应力、低温低应力这四种不同的情况，而在这四种外部条件下蠕变的机理存在一定的差异，第二，这 276 条样本中包含了不同体系(包含了 1 2 3 4 代)的合金，而不同体系合金的蠕变机理也存在较大的差异。而在现有的小样本的情况下，不适合对所有体系的合金统一建模。

针对上述的问题，基于聚类的最优回归集成学习方法首先依据合金的宏观、微观多尺度的特性以及外部环境因素，采用聚类将其聚成不同的合金类，然后基于聚类划分的结果，针对不同性能的合金类以预测精度为学习目标采用集成学习方法为每类合金构建适合的模型。

而通过实验验证，如图 6 所示，聚类方法区分出了不同体系的合金和不同的外部条件。



	Re_wt	Co_wt	Al_wt	Ti_wt	W_wt	Mo_wt	Cr_wt	Ta_wt	C_wt	B_wt
聚类中心1	3.382979	7.5	6.302128	0.510638	3.468085	1.414894	5.765957	6.840426	0.004	0.05
聚类中心2	1.847353	1.664412	5.054118	0.045588	7.536765	0.937059	7.216765	8.56	0.067765	0.026471
聚类中心3	1.505	5.9925	5.545	0.44	6.77	0.8375	7.76	7.1225	3.05E-17	-2.8E-18
聚类中心4	6.3	5.133333	5.338333	8.33E-16	6	-1.2E-16	2.8	8.866667	2.44E-17	-5.6E-18
聚类中心5	-4.2E-16	3.571429	5	1.642857	6.285714	-1.2E-16	9.714286	8.571429	0	0
聚类中心6	0	10	5	2	12	0	9	1.433333	0.103333	0.015
聚类中心7	1.25E-15	5.977778	4.211111	3.507222	6.258333	1.226111	9.519444	4.578889	0.028611	0.001644
聚类中心8	3.353191	8.17234	6.003723	0.910638	4.38617	1.180851	7.202128	5.948936	0.043883	0.004165

	Nb_wt	Hf_wt	S_s	1_a_s	2_a_ts	S_t	1_a_s	2_a_s	t	stress
聚类中心1	1.26E-16	0.179787	2	4	16	1307	1080	898	1022.021	266.6809
聚类中心2	1.078235	0.252353	4	4	20	1275	1080	871	1040	137.2
聚类中心3	5.39E-17	0.6225	3.85	4.2	21.8	1266.3	1088.95	828.675	1205.225	224.6415
聚类中心4	0.283333	0.116667	4	4	20	1180	1140	870	1050	193.3333
聚类中心5	-1.8E-17	3.89E-17	4.571429	4	32	2350	1975	1600	1657.143	387.3857
聚类中心6	0	2	3.333333	4	32	1536	1377.333	1113.333	1256.5	262.2
聚类中心7	1.8E-17	0.1325	2.472222	4	17.88889	1273.861	1079.472	885.7778	836.9722	390.0417
聚类中心8	3.23E-16	0.159574	2	4	17.02128	1307.766	1096.596	890.8511	977.3191	288.4894

图 6 聚类中心

其次，如图 7 所示，对比单一模型预测结果，基于聚类的最优回归集成学习方法能够更加准确的预测蠕变断裂寿命。

误差对比

	cluster_0	cluster_1	cluster_2	cluster_3	cluster_4	cluster_5	cluster_6	cluster_7	noCluster
样本数量	85	34	44	9	20	26	36	22	276
随机森林	0.0618228	0.036327	0.062012	0.122721	0.047645	0.09534	0.031056	0.158204	0.076207
支持向量	0.0768798	0.036327	0.066191	0.143715	0.06084	0.089363	0.049515	0.149752	0.087113
高斯过程	0.0685664	0.031109	0.067761	0.13199	0.050945	0.096139	0.085567	0.153316	0.078415

	MAE
Cluster	0.064618
noCluster	0.076207

$MAE = \frac{1}{N}\sum \frac{|x_{real} - x_{predict}|}{x_{real}}$ $RE = \frac{|x_{real} - x_{predict}|}{x_{real}}$ $AE = |x_{real} - x_{predict}|$

图 7 基于聚类的最优回归集成学习模型和单一预测模型的预测精度对比

如图 8 所示，虽然总体的预测精度提高，但还存在部分样本的预测误差相比未聚类方法更高。

部分样本聚类结果(MAE=0.076)					部分样本非聚类结果(MAE=0.088)			
聚类类别	Real	Predict_c	AE_c	RE_c	Real	Predict_n	AE_n	RE_n
0	4.70048	4.548324	0.152157	0.03237	4.70048	4.428126	0.272355	0.057942
0	4.728272	4.467424	0.260849	0.055168	4.728272	4.619413	0.10886	0.023023
2	4.730039	4.329928	0.400112	0.084589	4.730039	4.327803	0.402236	0.085039
2	4.738827	4.82638	0.08755	0.018476	4.738827	4.388802	0.350025	0.073863
5	4.740575	4.254225	0.48635	0.102593	4.740575	5.652942	0.91237	0.192459
2	4.749271	4.391292	0.357978	0.075375	4.749271	4.694555	0.054716	0.011521
0	4.779963	4.435264	0.344699	0.072113	4.779963	4.417884	0.36208	0.075749
0	4.780803	4.404914	0.375889	0.078625	4.780803	4.380474	0.400328	0.083737
0	4.785824	4.301287	0.484536	0.101244	4.785824	4.563072	0.222751	0.046544
2	4.799091	4.351214	0.447877	0.093325	4.799091	4.235677	0.563414	0.1174
0	4.840242	4.026177	0.814065	0.168187	4.840242	4.247004	0.593238	0.122564
5	4.896346	5.28498	0.38863	0.079372	4.896346	4.30728	0.589066	0.120307
2	4.964242	4.803477	0.160765	0.032385	4.964242	4.235865	0.728377	0.146725

图 8 预测结果对比

2.3.2 基于主动学习的多层级交互式特征分析方法

目前的特征选择算法选择特征时存在不稳定性，可能会剔除专家认为很关键的特征给剔除掉，这一方面可能会影响材料领域专家对新属性的计算；另一方面也可能降低机器模型的预测精度。因此在进行特征选择时，需要综合考虑领域专家经验、机器学习模型的预测精度等因素，协同完成特征分析。因此课题组针对上述问题，提出并完成了基于主动学习多层级交互式特征分析方法，流程如图 9 所示。针对高温合金数据可能存在的稀疏性、冗余性、不相关性、高维度等问题展开了逐层的特征分析，初步设计了与实现了专家经验的表示和融入方法，特征分析的阈值确定以及算法筛选条件的学习，模型选择和多目标评价函数的确定，并设计了特征重要性集成方法。利用该方法可以定性定量分析各种因素对高温合金性能的影响程度。

多层级交互式特征分析方法集成了多领域专家的知识和经验，并对特征分析得到的结果采用机器学习模型进行验证。算法、模型验证、领域专家经验三者共

同保证筛选特征子集的质量。该方法在获得最优特征子集的同时，也能有效的分析属性与属性之间的关联关系，属性与性能之间的因果关系。

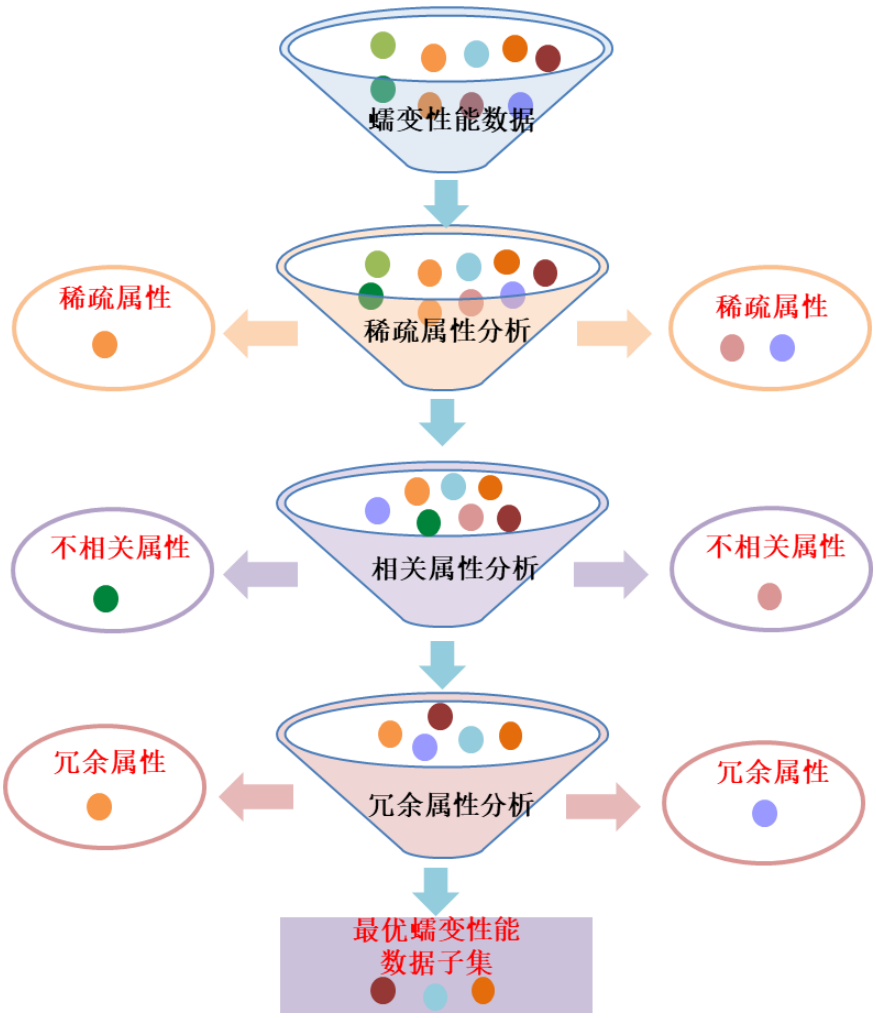


图 9 多层级特征分析方法流程

1) 蠕变新属性计算

基于上述的预测蠕变断裂寿命的结果发现单纯的从改进机器学习模型的角度以更好的描述蠕变性能(提升预测精度)是不够的，样本的质与量也是影响建模效果的关键因素之一。现有的样本中描述蠕变的主要是宏观方面的成分、热处理和外部条件。影响蠕变的还有很多微、介观因素(如扩散、晶格常数和层错能等)。对这些因素的定性定量刻画是关键。

王院士在 49 篇文献中从计算的角度研究镍基单晶高温合金的占位、位错扩散等，这些都是影响蠕变的重要因素，而韩国的 Young-Kwang Kim 等人通过整合前人的工作得出了计算蠕变的解析式并提供了大部分相关的微观属性的计算

公式，部分公式如图 10 所示，其中包含了晶格常数、层错能、扩散系数和剪切模量等在 49 篇文献中出现多次的物理量，并且给出了相关的计算方法。使得我们可以通过已有的宏观数计算晶格常数等微观描述因子，解决了缺少微观描述因子的问题。

$$\dot{\varepsilon}_m = \dot{\varepsilon}_m^{\gamma\ matrix} + \dot{\varepsilon}_m^{GBS} = A_1\phi(1-\phi)\left(1/\phi^{1/3}-1\right)\frac{D_L G b}{kT}\left(\frac{\bar{\lambda}}{b}\right)^2\left(\frac{\Gamma}{Gb}\right)^3\left(\frac{\sigma-\sigma_p}{G}\right)^5 + A_2\frac{D_L G b}{kT}\left(\frac{b}{d}\right)^2\left(\frac{\sigma}{G}\right)^2$$

图 10 最小蠕变速率公式

具体的计算过程如图 11 所示，首先，把已有的宏观数据(包含了成分、外部条件)作为 Thermo Calc 热力学计算的输入，计算出对应的相体积分数、元素占位和元素分配，结合 Young-Kwang Kim 等人给出的计算方法计算出新的微观、介观属性------扩散系数、层错能、晶格常数和剪切模量。最后结合原有的宏观属性(成分、热处理、外部条件)、计算的中间结果(相体积分数)和新计算的微观、介观属性(扩散系数、层错能、晶格常数和剪切模量)构建了跨尺度的学习样本。

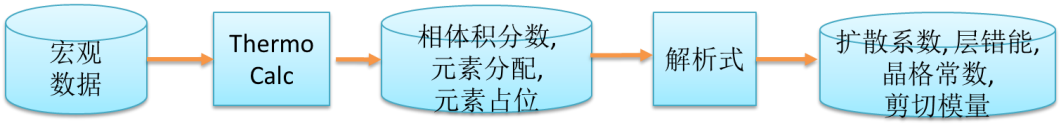


图 11 新属性计算过程

最后结合原有的宏观属性(成分、热处理、外部条件)、计算的中间结果(相体积分数)和新计算的微观、介观属性(扩散系数、层错能、晶格常数和剪切模量)构建了跨尺度的学习样本，如图 12 所示。

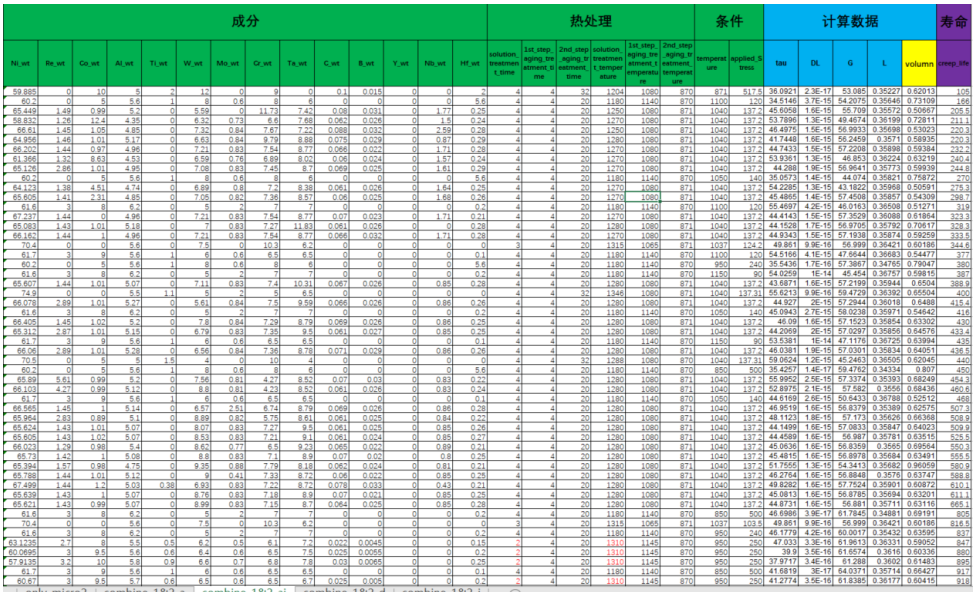


图 12 多尺度学习样本

为了验证计算新属性的有效性，我组对比了只有已有的宏观属性数据的样本、只有计算属性数据的样本和结合二者的样本在相同的机器学习模型上做蠕变预测的效果。实验结果如表 1 所示，相比于只有宏观属性的样本，拥有多尺度属性的样本预测蠕变寿命的精度更低，原因可能是相对于小样本，维度更高，导致得精度下降，所以有必要进行特征选择

表 1 不同样本的预测 MAE

	宏观属性	微观属性	多尺度属性
随机森林	0.043653	0.08450536	0.04407963
SVR	0.048518	0.09196395	0.04922684
高斯过程	0.032687	0.0827986	0.04120816

2) 特征分析方法在蠕变数据中的应用

为了验证我们提出的特征分析方法的有效性和可行性，我们将计算好蠕变新属性的 78 条蠕变性能数据作为学习样本，样本成分维度 14 维，包括 Ni, Re, Co, Al, Ti, W, Mo, Cr, Ta, C, B, Y, Nb, Hf 各元素的质量分数，热处理维度 6 维，包括固溶温度、固溶时间、第一阶段时效处理温度、第二阶段时效处理温度、第一阶段时效处理时间、第二阶段时效处理时间，外部条件维度 2 维，包括外部温度、外部应力，目标属性（蠕变断裂寿命）1 维，计算新属性 5 维，包括相摩尔分数、层错能、晶格常数、剪切模量、扩散系数。

首先，我们对样本身数量为 78 条，描述因子有 27 个的蠕变性能数据进行稀疏属性分析。稀疏属性指的是当属性值为离散值时比如 0,1，如果其中某个离散值的数量超过总数量的 95%，说明该离散值对应的属性为稀疏属性；当属性值为连续值时，如果属性值的方差小于给定阈值时，也说明该属性为稀疏属性。通过稀疏属性分析可以过滤一些稀疏属性，使得各个属性值的分布尽可能均匀，从而为后续的数据分析提供良好质量的数据，提高机器学习的预测精度。如图 13 所示，根据专家经验，稀疏值阈值设定为 95，从左图可以看出 Y（钇）元素的质量分数和 2sat（二阶段时效处理时间）这两个离散属性的稀疏值大于 95，因此它们都是稀疏属性。依据专家经验，属性方差阈值设定为 0.01，从右图中可看出所有连续属性方差都大于 0.01，因此保留全部的连续属性。稀疏属性分析后，保留了

25 个属性。

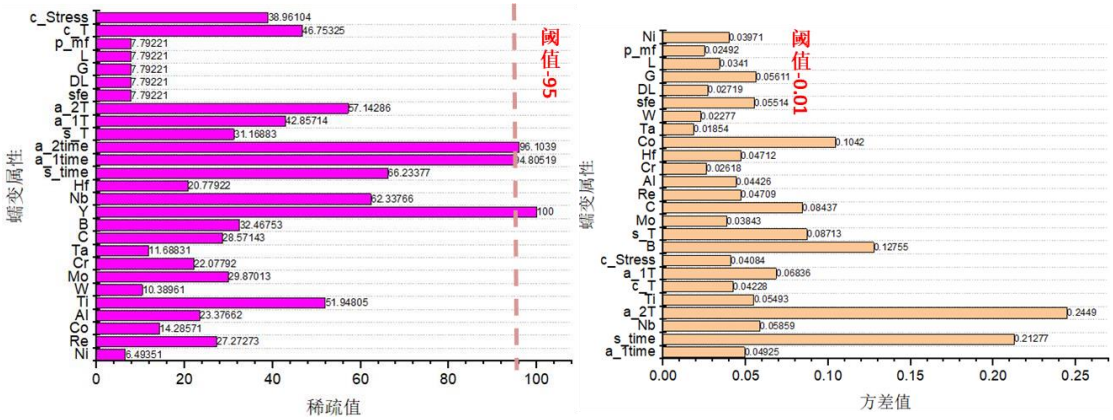


图 13 稀疏属性分析结果

其次，蠕变性能数据在经过稀疏属性分析后，虽然消除了特征集中存在的稀疏信息，但特征集中仍然可能存在不相关信息、冗余信息。因而我们对蠕变数据做了相关属性分析。相关属性分析可以过滤无关或弱相关的属性，保留最相关的属性，使得条件属性与决策属性之间的相关性较强，从而为后续的数据分析提供良好质量的数据，提高机器学习的预测精度。从表 2 中可以看出，固溶处理中，s_time(固溶处理时间)的相关系数最高为 0.63, s_T(固溶温度)的相关系数为 0.36。元素中，相关系数比较高的是 Co(0.43) Nb(0.34) Al(0.3) Re(0.3)。微观属性中，剪切模量 G 的相关系数为 0.4，层错能 sfe 为 0.32，扩散系数 DL 为 0.29,晶格常数 L 为较低的 0.08 这里和蠕变关系比较大的外部应力 c_Stress 的相关系数为 0.04，因为 78 条样本中，大部分的应力是相同的。依据专家经验，相关系数阈值设定为 0.02，从左图可以看出，所有的蠕变属性（25 个）与决策属性（蠕变断裂寿命）之间的相关系数都大于 0.02，因此保留了全部属性。

表 2 蠕变属性与蠕变断裂寿命之间的相关系数表

描述因子	s_time	Co	a_2T	G	c_T	Ni	s_T
相关度	0.633935	0.436919	0.42015	0.405154	0.395389	0.380697	0.366304
描述因子	Nb	a_1T	sfe	B	Al	Re	DL
相关度	0.349064	0.331214	0.316103	0.30762	0.303095	0.300944	0.290518
描述因子	Ti	Cr	C	Mo	Hf	p_mf	L
相关度	0.236486	0.23151	0.202839	0.159675	0.138903	0.08824	0.084786
描述因子	W	Ta	c_Stress	a_1time			
相关度	0.075546	0.058569	0.043993	0.02473			

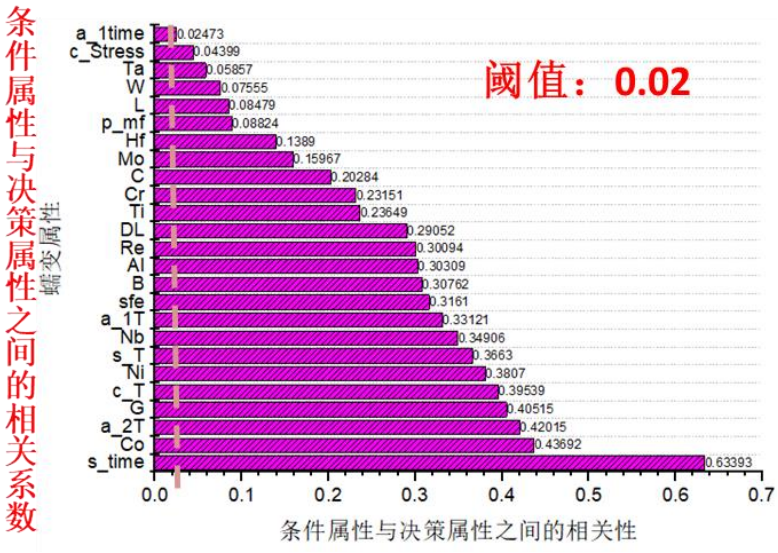


图 14 相关属性分析

最后，蠕变属性与属性之间也可能存在较强的相关性，称为冗余性。我们通过冗余属性分析可以去除属性与属性之间的冗余信息。如图 14 所示，观察此图可以得知，图中给出了蠕变属性与属性之间的相关系数，相关系数的绝对值越大，表明两者之间相关程度越大，两属性之间存在冗余性。从分析结果中可以看出，C 和 B 元素质量分数之间的相关系数为 0.935285，存在较强的相关性，因此可以认为二者之间可能存在冗余性。a_2T（二阶段时效处理温度）和 B 元素之间的相关关系为 0.869216 ，相关性较强，因此二者之间也可能存在冗余性。

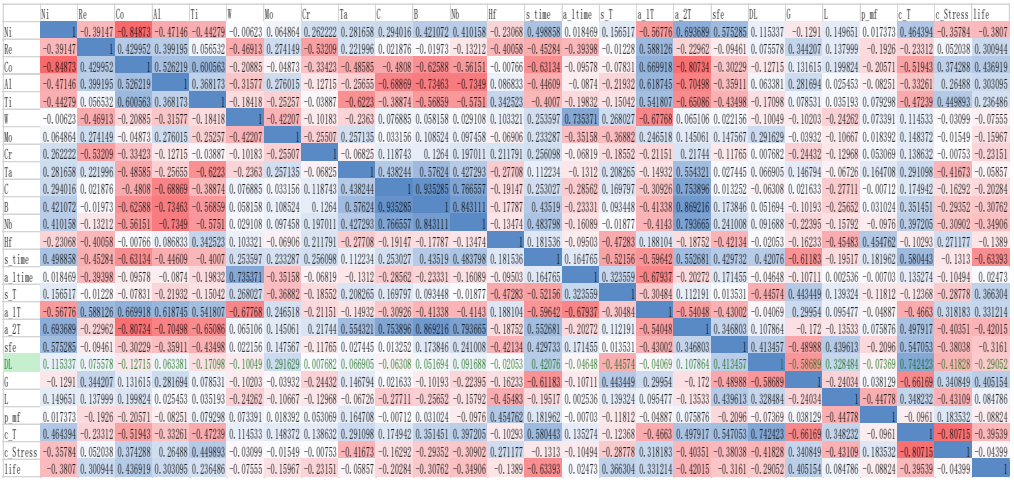


图 15 冗余属性分析

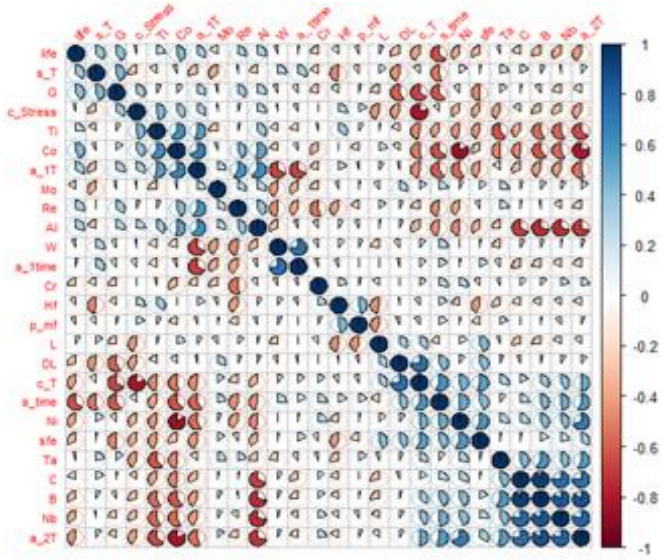


图 16 蠕变属性与蠕变属性之间的相关系数图

经过多层级特征分析方法过滤和筛选后，Al（铝）、Y（钇）元素的质量分数、二阶段时效处理温度、二阶段时效处理时间四个属性被筛除。对比筛选过特征的样本和未筛选特征的样本在同一机器学习模型上做蠕变预测，结果如表 3 所示。

表 3 特征筛选结果

机器学习方法	多尺度未筛选	多尺度筛选
随机森林	0.04408	0.04298
SVR	0.04923	0.04920
高斯过程	0.04114	0.04025

结果显示经过特征筛选之后，预测精度确实提高了。证明了方法的有效性。但提升幅度不明显，主要是因为样本少，未来的工作之一是计算更多的样本数据的数据。

2.4 平台研发和建设

我组至项目启动开始就一直在搭建高温合金机器学习演示平台，以方便展示本课题的研究成果。首先我组完成了平台总体架构的设计，如图 17 所示，为高温合平台的研发和建设提供依据和指导。其次，我组基于设计好的平台架构图，已经基本实现高温合金机器学习平台，该平台完成了材料实验数据与计算数据、材料计算方法与面向材料的机器学习 / 数据挖掘方法的上传、下载，上传的数据与数据之间存在关联关系，数据能够溯源，并且数据上传后能够对数据进行质量检查并可以提醒用户进行修正；该平台为每一个使用用户构建了用户画像（模型），

用户模型定义为 User_model = U{<用户基本信息(姓名，职位，所在单位，联系方式...)>，<用户行为偏好(时间偏好，作者偏好，文献偏好，机构偏好)>，<用户等级(计算机等级、材料等级)>，...}，如图 18 所示，从而实现了平台搜索功能的个性化，将最可能满足用户需求的搜索结果优先展示给用户；该平台设计和实现了多层级交互式特征分析方法，该方法能够定性定量刻画高温合金属性与性能之间的关联关系和因果关系，交互性强，并以可视化的形式展现给用户。最后，该平台为了实现对镍基高温合金性能的预测，第一，完成了非自动化的机器学习算法，用户选择数据、算法，系统执行算法并展示实验结果，如图 19 所示；第二，设计和实现了半自动化的机器学习算法流水线，为用户提供算法模板，根据用户选择的模板类型为用户自动生成相应的模板流水线，在用户选择需要处理的科学数据后，对该数据进行自动化处理，如图 20 所示；第三初步探索和研究了数据驱动的全自动式的机器算法，如图 21 所示。

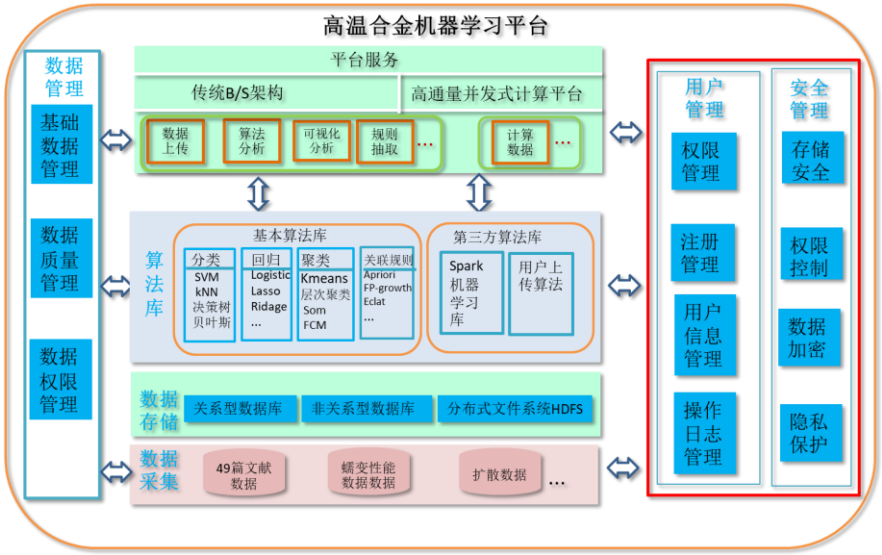


图 17 平台总体架构

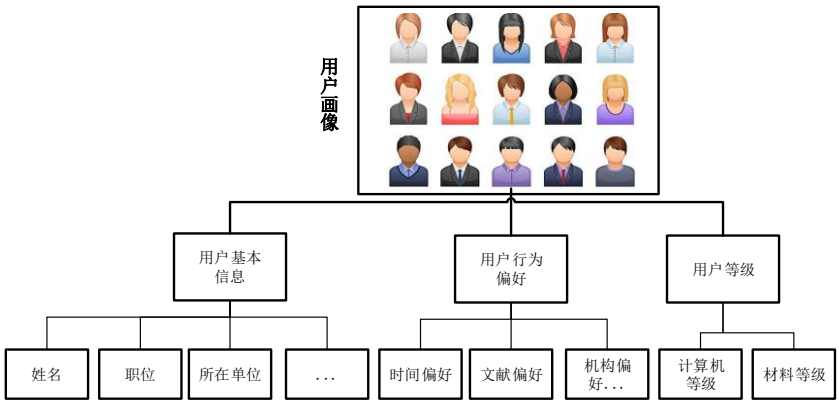


图 18 用户模型

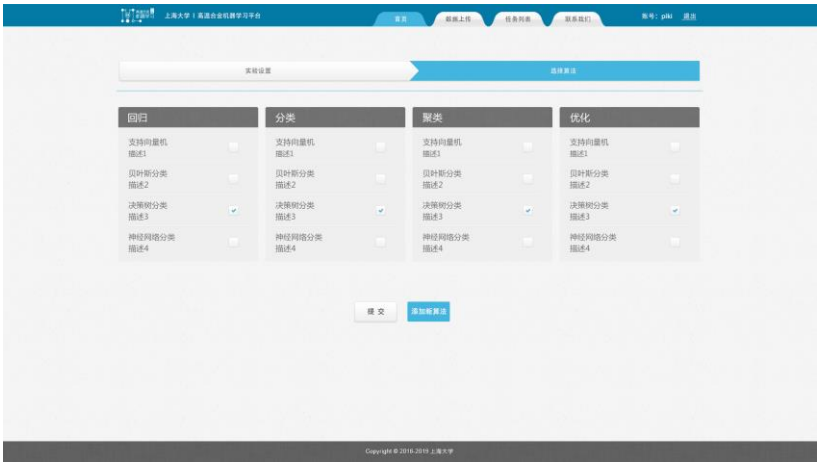


图 19 非自动化机器学习算法界面

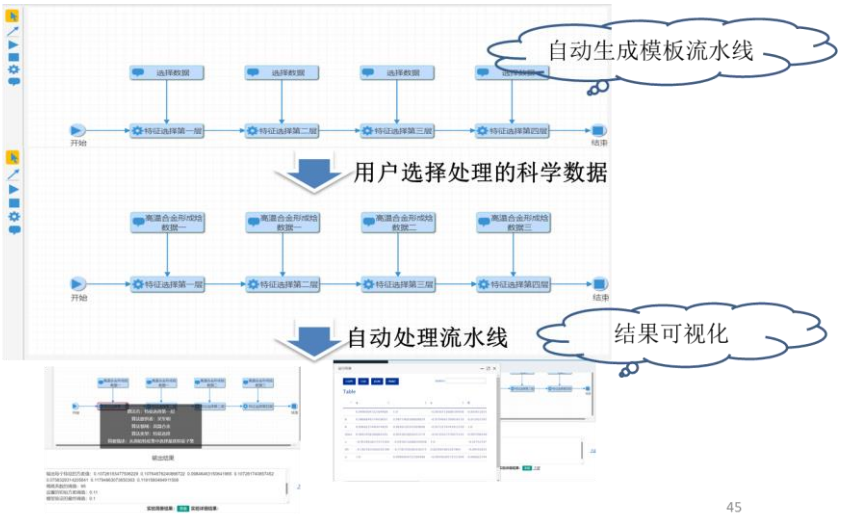


图 20 半自动式机器学习算法流水线

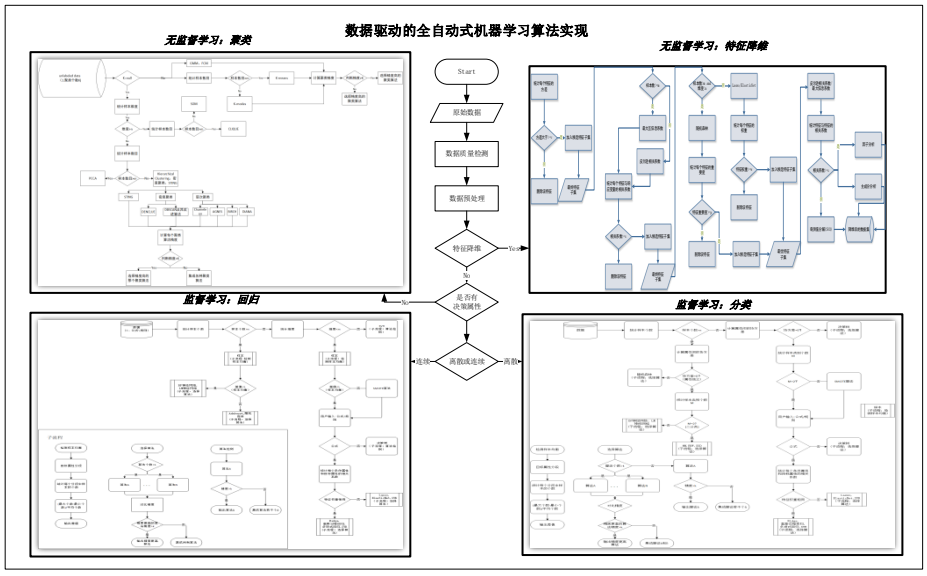


图 21 数据驱动的全自动机器学习算法

3. 存在问题与展望

就目前来看，我组工作中存在的问题主要包括三方面：**数据的质和量问题并存，如何分析高通量计算产生的结果数据，如何解释机器学习结果**。

问题一：数据的质与量并存

- **国际高温合金会议文献数据**

100 篇文献中，大部分都是聚焦于镍基单晶高温合金，但其研究的合金体系繁多，探索的是不同成分（多组元）、结构（合金相组成、错配位错、位错运动等）、外界条件（初熔温度、固溶处理、时效处理等）与合金持久性能（蠕变性能、疲劳断裂寿命、力学性能等）的关联关系。但由于合金数据的多体系化，导致数据的质难以得到保证。如何解决数据的量与质之间的不均衡关系迫切需要领域专家经验的融入。

文献中的数据包括了计算数据和实验数据，主要是针对不同合金体系进行计算和实验得到的数据。如何实现计算数据和实验数据的整合也是我们面临的主要问题。

文献中合金性能的描述因子众多，不同文献研究了不同因素对合金性能的影响。在构造学习样本时会出现大量的空缺属性，需要领域专家或通过材料计算或实验来补充。

- **蠕变数据**

收集的蠕变数据主要来源于不同的文献或专利；涵盖多体系的高温合金，研究的合金属性各不相同；有些数据通过对文献或专利中图片数据利用描点工具转换而得，数据的正确性和完整性较差。因而需要专家经验的引入，对问题数据进行解释；对缺失数据进行补充；课题组材料专家通过计算或实验得到可学习的数据；通过文献/专利等来收集和完善数据。

问题二：如何分析高通量计算产生的结果数据

机器学习驱动的高通量计算：分析过程中有大量的合金属性需要通过计算获得。高通量计算结果驱动机器学习：高通量计算机得到的结果是否可以学习？如何学习？

问题三：如何解释机器学习结果

针对已收集的高温合金数据和蠕变数据，利用多项式拟合、最小二乘拟合、

指数拟合、对数拟合等机器学习回归方法进行学习建模，从而获得合金性能描述因子与性能之间的参量解析式 P =F（X）;利用机器学习方法如决策树、随机森林等对数据进行规则抽取，构建适用于高温合金数据的规则库。

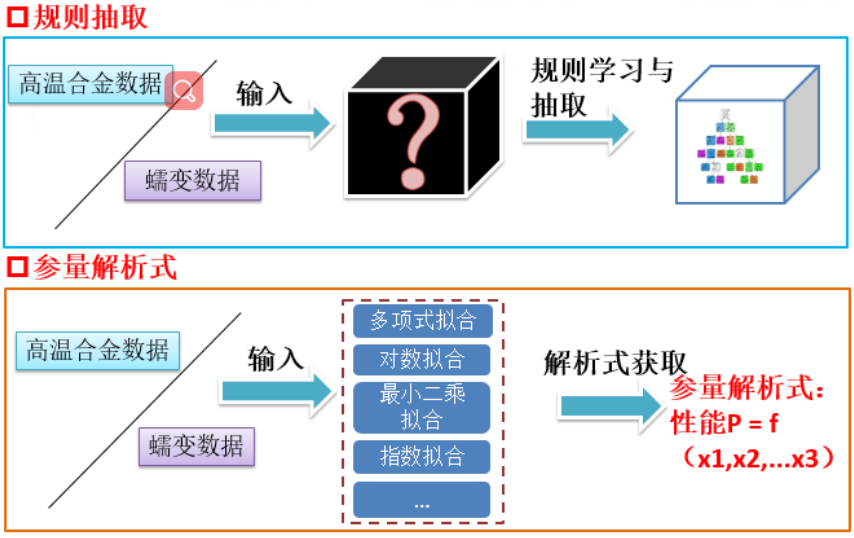


图 22 规则抽取和参量解析式获取

4. 下一步工作计划

我组在接下来的工作中将进一步与领域专家确定学习目标与研究问题，从数据、方法、结果三个方面展开工作。**数据方面**，通过文献/专利、计算和实验获得更多的数据，提高数据的质和量，从而保证机器学习/数据挖掘方法分析结果的正确性和可信性；**方法方面**，在数据驱动下，研究面向高温合金的机器学习/数据挖掘方法，并将其分析结果与领域专家的研究成果相结合，不断改进和完善学习模型，从而提高分析结果的正确性和实用性；**结果方面**，采用规则抽取/数据可视化等方法与技术对分析结果进行再处理，并和领域研究成果相结合，提高分析结果的可理解性和可用性。