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，在过去的半年时间里，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。

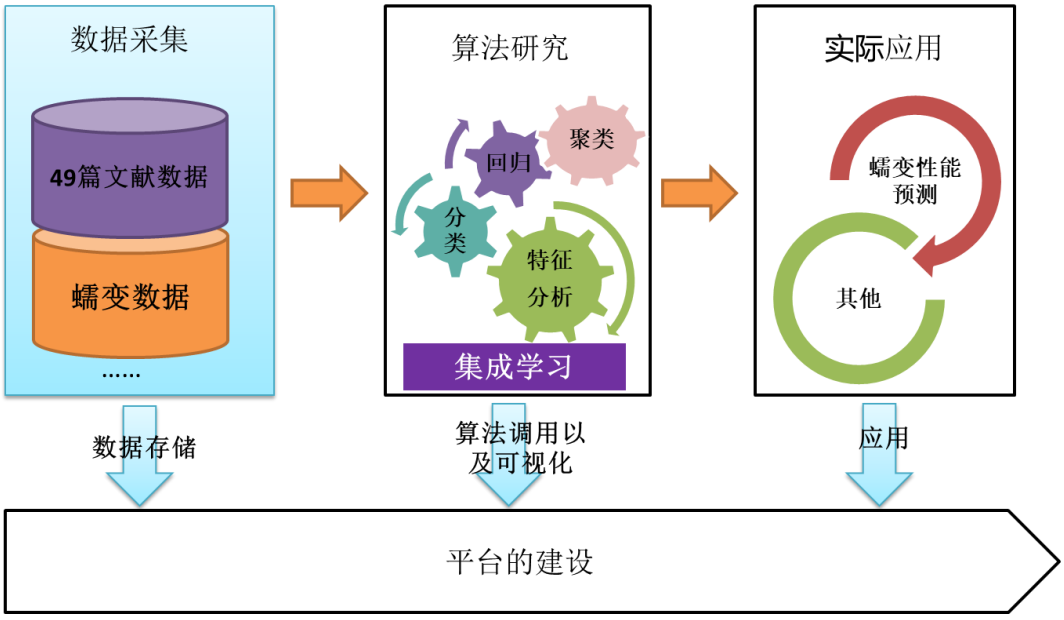

图 1 最新工作总览

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。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。

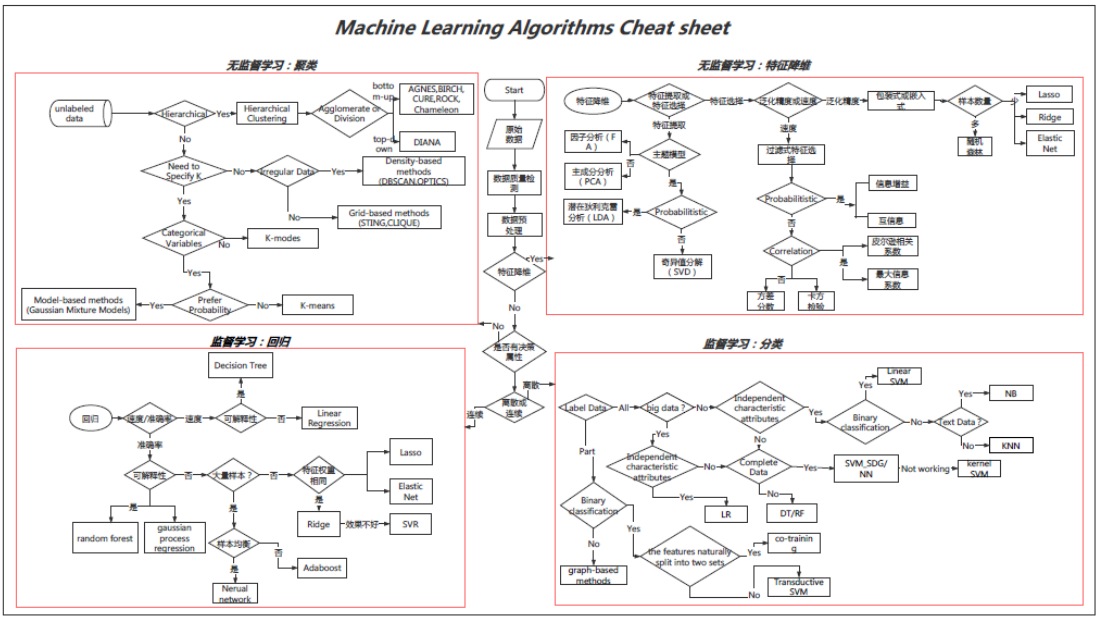

图 2 机器学习算法分类树

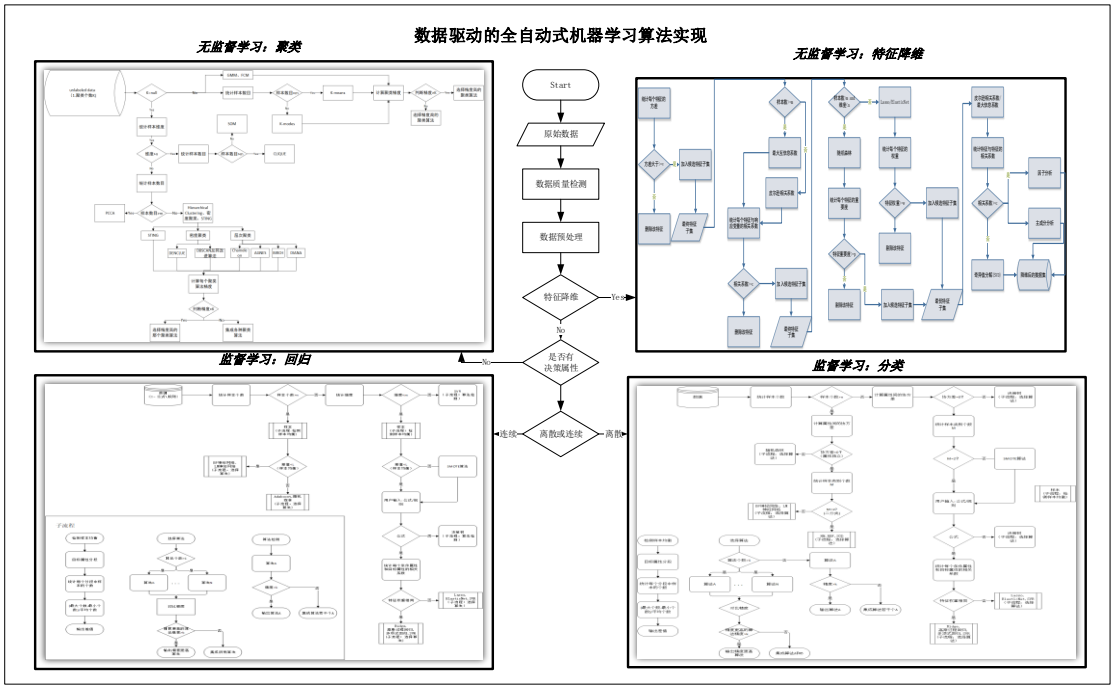

图 3 数据驱动的全自动式机器学习算法流程图·

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

2) 在蠕变数据中的应用

2.3.1 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 4 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。

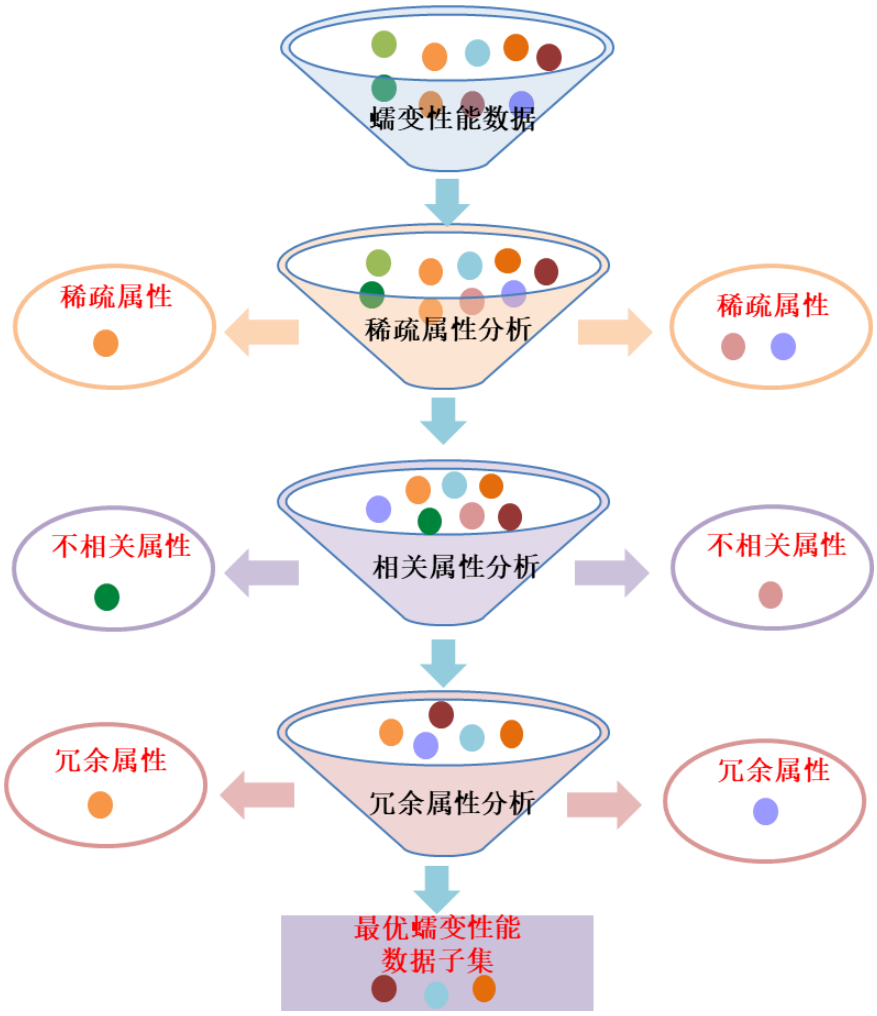

图 4 多层级特征分析方法流程

1) 蠕变新属性计算

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 5 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。

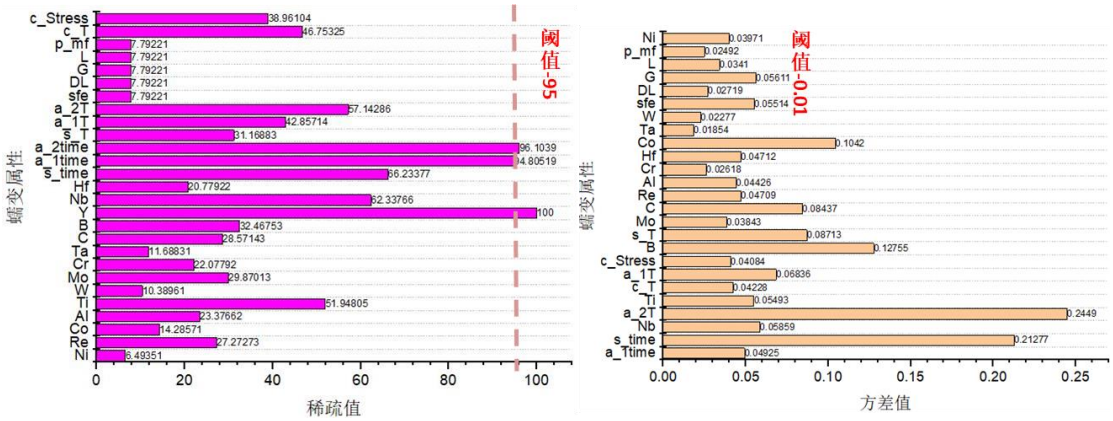

图 5 稀疏属性分析结果

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 1 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 1 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			

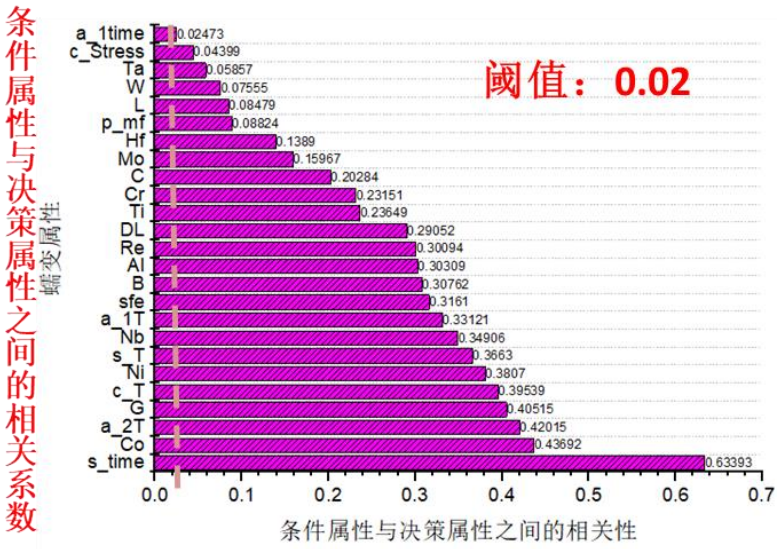

图 6 相关属性分析

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 7 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。

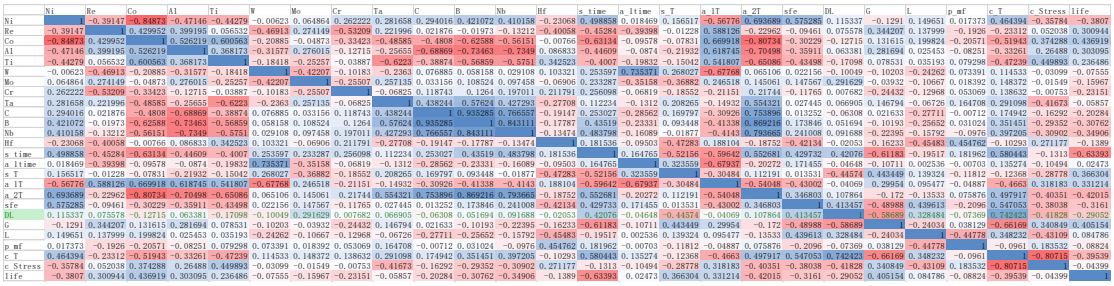

图 7 冗余属性分析

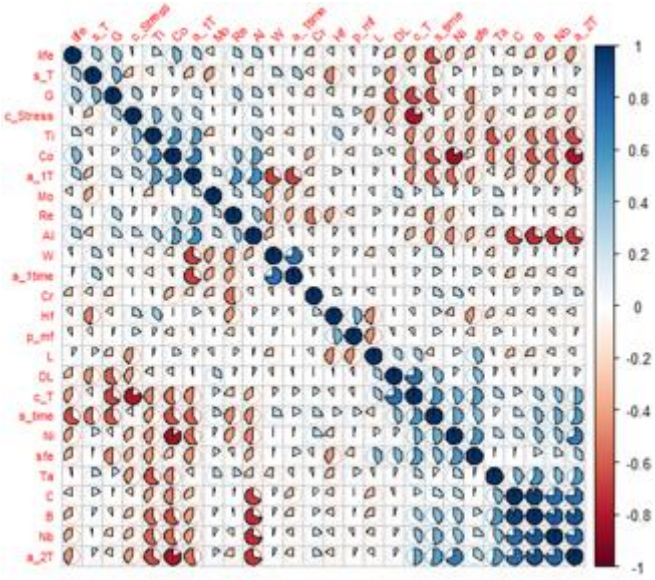

图 8 蠕变属性与蠕变属性之间的相关系数图

2.4 平台研发和建设

我组正在搭建高温合金机器学习演示平台，方便展示本课题的研究成果。该平台已初步完成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、个性化搜索，并设计和实现了多层级交互式特征分析方法。

3. 总结与展望