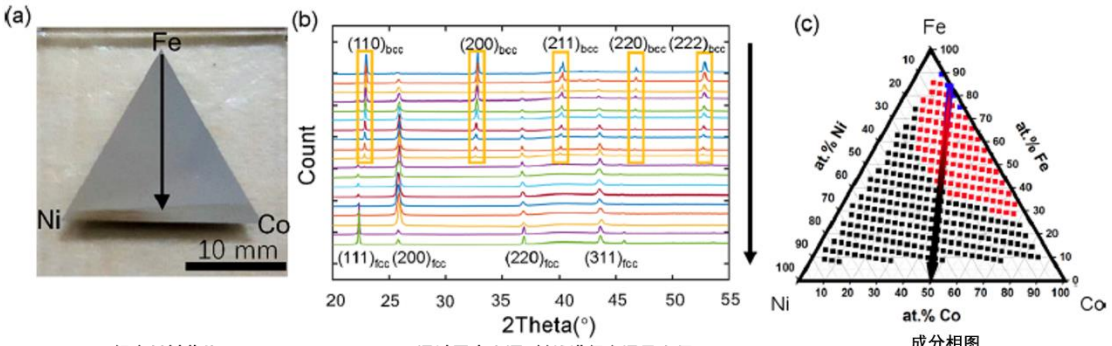
通过组合材料芯片技术快速构建 Fe-Co-Ni 成分相图

班级：国科博 17 班

学号： B20170427

姓名： 赵朝阳



组合材料芯片　　通过同步光源X射线进行高通量表征　　成分相图（机器学习）

摘要：

通过磁控共溅射制备 100nm 厚的 Fe-Co-Ni 材料芯片并在 500，600，700℃下绝热退火。逐点的成分和结构匹配由同步光源的微束 X 射线表征。衍射谱的图像按照每秒一张的速率被记录下来。XRD 衍射图谱被自动处理，相的识别与归类由层次聚类的算法完成，并用来构建组合成分相图。所构建的相图与 ASM 合金相图数据库记载的绝热章节内容一致，验证了本文构建相图方法的有效性。

关键词：组合材料芯片、Fe-Co-Ni，X 射线衍射，层次聚类，相图

构建相图的传统方法是每次对一个样品进行特征提取和表征，此法耗时，没有系统的流程，在科技飞速发展的现今远不能满足科研人员的需求。自上世纪 60 年代起，进行过许多匹配成分-相关系的尝试。组合材料技术、特征高通量提取、组合材料库的快速表征展示出加速材料筛选和优化的巨大潜力。Yoo 等人利用 Fe-NI-Co 三元合金系统说明了使用包含连续成分扩散的组合材料芯片薄膜来建立晶体结构和成分之间的复杂关系是可行的。JangHorban 等人展示相似类型的样品，Cr-Ni-Re 在 1100℃下的相图与已发表文献中的相图一致。

通过 XRD 构建组合材料库的通量很大程度上受限于 X 射线的流量和射线束斑的大小。同步衍射为高空间分辨率的快速表征提供了一个理想的 X 射线源。进一步的加速可通过使用带有区域探针的聚焦 X 射线微束实现，其避免了耗时的角度扫描。经证实，组合材料库的衍射谱可通过 1~30s 的同步衍射获取。

为了适应高通量衍射实验产生的大量数据，需要自动化相的识别和和聚类的流程。在衍射谱的数据处理工作中用到了机器学习的方法。举个例子，Bunn 等人使用 XRD、拉曼荧光光谱和成分数据分析 Ni-Al 薄膜的相形成和氧化物时采用 Adaboosting 特征学习，一种监督式机器学习方法。该机器方法的方法的一大弊端是需要大量的训练数据，这在实际中通常是不可行的。Long 等人使用非负矩阵分解(NMF)，一种非监督式机器学习方法，用于相的匹配以减少对数据量的需求，对基本的物理意义考虑的较少。许多微分方法，如 CombiFD、GRENDEL 和 AgileFD 被开发用来确保结论模型携带足够的物理含义。由于 XRD 峰位的偏移会覆盖较大的成分范围，直到目前，消除峰位偏移仍是对相匹配算法的一大挑战。

聚类是一项根据特定测量将数据分组的技术。在最近的工作中，Iwasaki 等人利用取自材料组合数据库的 X 射线衍射数据结合层次聚类，通过尝试不同的相似度度量方法以比较每种相似度度量方法的有效性。常见的相似度度量方式有以下几种：

①皮尔逊相关系数(Pearson Correlation Coefficient)

皮尔逊相关系数一般用于计算两个定距变量间联系的紧密程度，它的取值在[-1，+1]之间。

$$\mathbf{p}(\boldsymbol{x},\boldsymbol{y})=\frac{\sum \boldsymbol{x_i y_i}-\boldsymbol{n\bar{x}\bar{y}}}{(\boldsymbol{n}-\mathbf{1})\boldsymbol{S_x S_y}}$$

$\boldsymbol{S_x},\boldsymbol{S_y}$是 x 和 y 的标准偏差

原理：用来反应两个变量线性相关程度的统计量

范围：[-1，+1]，绝对值越大，说明相关性越强，负相关对于推荐的小。

该相似度并不是最好的选择，也不是最坏的选择，只是因为其容易理解，在早期研究中经常被提起。使用 Pearson 线性相关系数必须假设数据是成对的从正态分布中取得的，并且数据至少在逻辑范畴内必须是等间距的数据。Mahout 中，为皮尔逊相关计算提供了一个扩展，通过增加一个枚举类型(Weighting)的参数来使得重叠数也成为计算相似度的影响因子。

②欧几里得距离(Euclidean Distance)

最初用于计算欧几里得空间中两个点的距离，假设 x，y 是 n 维空间的两个点，它们之间的欧几里得距离是：

$$\mathbf{d}(\boldsymbol{x},\boldsymbol{y})=\sqrt{\sum(\boldsymbol{x_i}-\boldsymbol{y_i})^2}$$

可以看出，当 n=2 时，欧几里得距离就是平面上两个点的距离。当用欧几里得距离表示相似度，一般采用以下公式进行转换：距离越小，相似度越大。

$$\mathbf{sim}(\boldsymbol{x},\boldsymbol{y})=\frac{\mathbf{1}}{\mathbf{1}+\boldsymbol{d}(\boldsymbol{x},\boldsymbol{y})}$$

原理：利用欧式距离 d 定义的相似度 S，S=$\frac{\mathbf{1}}{\mathbf{1}+\boldsymbol{d}}$

范围：[0，1]，值越大，说明 d 越小，距离越近，相似度越高

说明：同皮尔逊相似度一样，该相似度也没有考虑重叠数对结果的影响，同样的，Mahout 通过增加一个枚举类型的参数来使得重叠数也成为计算相似度的影响因子。

③Cosine 相似度(Cosine Similarity)

Cosine 相似度被广泛用于计算文档数据的相似度：

$$\mathbf{T}(\boldsymbol{x},\boldsymbol{y}) = \frac{\sum \boldsymbol{x_i y_i}}{\sqrt{\sum \boldsymbol{x_i^2}}\sqrt{\sum \boldsymbol{y_i^2}}}$$

原理：多维空间两点与所设定的点形成夹角的余弦值。

范围：[-1，1]，值越大，说明夹角越大，两点的距离就越远，相似度就越小。

说明：在数学表达中，如果对两个项的属性进行数据中心化，计算出来的余弦相似度和皮尔逊相似度是一样的，在 Mahout 中，实现了数据中心化的过程，所以皮尔逊相似度也是数据中心化后的余弦相似度。

④Tanimoto 系数(Tanimoto Coefficient)

Tanimoto 系数也称为 Jaccard 系数，是 Cosine 相似度的扩展，也多用于计算文档数据的相似度：

$$\mathbf{T}(\boldsymbol{x},\boldsymbol{y}) = \frac{\sum \boldsymbol{x_i y_i}}{\sqrt{\sum \boldsymbol{x_i^2}} + \sqrt{\sum \boldsymbol{y_i^2}} - \sum \boldsymbol{x_i}\,\boldsymbol{y_i}}$$

原理：是对 Cosine 相似度的扩展。

范围：[0，1]，完全重叠时为 1，无重叠项时为 0，越接近 1 越相似。

说明：适合处理无偏好的打分数据。

⑤曼哈顿距离(Manhattan Distance)

曼哈顿距离公式为：

$$\mathbf{d}(\boldsymbol{X_i},\boldsymbol{X_j}) = \sum_{\boldsymbol{k=1}}^{\boldsymbol{n}} \left|\boldsymbol{x_{ik}} - \boldsymbol{x_{jk}}\right|$$

$\boldsymbol{X_i} = (\boldsymbol{x_{i1}}, \boldsymbol{x_{i2}}, \dots, \boldsymbol{x_{in}})^{\boldsymbol{T}}$，$\boldsymbol{X_j} = \left(\boldsymbol{x_{j1}}, \boldsymbol{x_{j2}}, \dots, \boldsymbol{x_{jn}}\right)^{\boldsymbol{T}}$为 n 维曼哈顿空间$\boldsymbol{R^n}$中的两个对象。

原理：同欧氏距离相似，均用于多维数据空间距离的测量。

范围：[0，1]，同欧氏距离一致，值越小，说明距离值越小，相似度越大。

说明：比欧氏距离计算量小，计算性能相对较高。

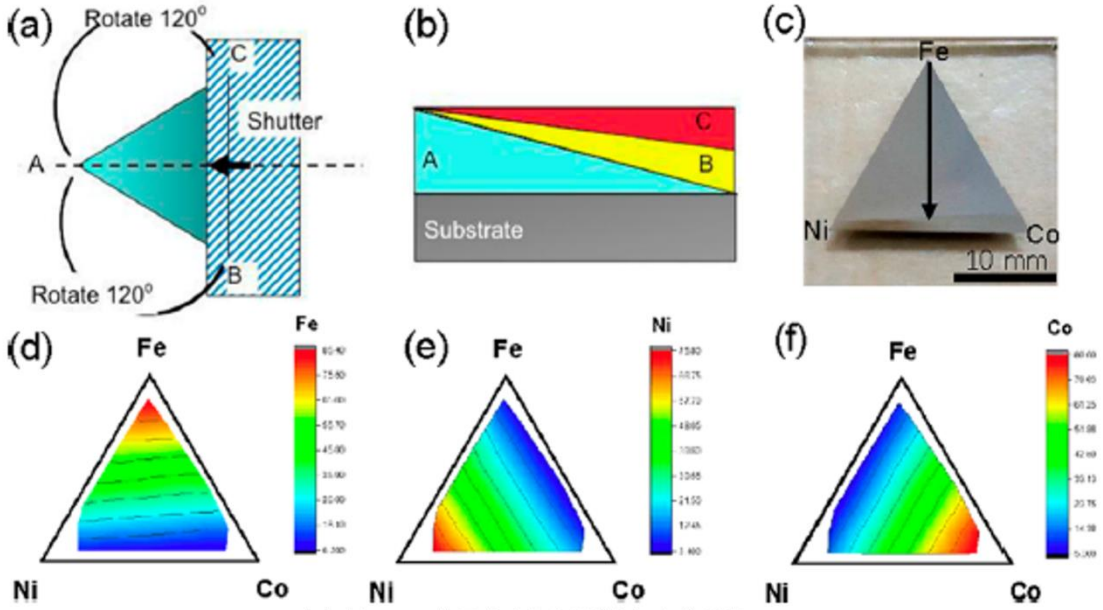
⑥切比雪夫距离公式(Chebyshev Distance)

切比雪夫距离公式为$\mathbf{d}(\boldsymbol{X_i},\boldsymbol{X_j}) = \max_{\boldsymbol{1 \ll K \ll n}} \left|\boldsymbol{x_{ik}} - \boldsymbol{x_{jk}}\right|$，其中，$\boldsymbol{X_i} = (\boldsymbol{x_{i1}}, \boldsymbol{x_{i2}}, \dots, \boldsymbol{x_{in}})^{\boldsymbol{T}}$，$\boldsymbol{X_j} = \left(\boldsymbol{x_{j1}}, \boldsymbol{x_{j2}}, \dots, \boldsymbol{x_{jn}}\right)^{\boldsymbol{T}}$为 n 维曼哈顿空间$\boldsymbol{R^n}$中的两个对象。

他们发现 Cosine、Pearson 相关系数和 Jensen-Shannon 离散度相似度测量技术，在有峰高变化和峰位的随机偏移的情况下，能呈现最好的结果。通过结合吉布斯相规则，Suram 等人使用 AgileFD-Gibbs 算法构建了具备物理含义的 V-Mn-Nb 氧化物相图。

在本文的工作中，一套快速构建成分相图的系统化的工作流得以建立。工作流中包括组合材料芯片的准备、使用 X 射线微束衍射和 X 射线荧光进行成分表征和基于层次聚类的自动化数据分析。为了对工作流进行演示，本文选择记录详实的 Fe-Co-Ni 三元合金系统。结果成分相图与 ASM 合金相图的绝热章节部分对比以便确认结果的正确性。

薄膜组合材料芯片能覆盖 Fe-Co-Ni 三元合金系统的整个成分变化范围。通过使用自行设计的高通量组合离子束沉积系统(HTC-IBD)，芯片被沉积在石英基底上。



（a）插图显示使用移动遮光器进行沉积的过程

（b）具备渐变成分扩散的多层薄膜的交叉区域

（c）制备的Fe-Co-Ni等边三角形成分扩散样品

（d），（e），（f）分别表示Fe、Ni、Co的成分扩散分布

图一

芯片上每点的成分由该点的厚度确定

$$C_i = \frac{t_i\rho_i/Z_i}{\sum_i t_i\rho_i/Z_i}, \sum_i t_i = const$$

t 是样品的厚度，ρ是密度，Z 是每个元素的原子质量。

100nm 厚的多层薄膜密封在抽真空的石英管中，分别在 500℃、600℃、700℃下绝热处理 2 小时，随后空冷。

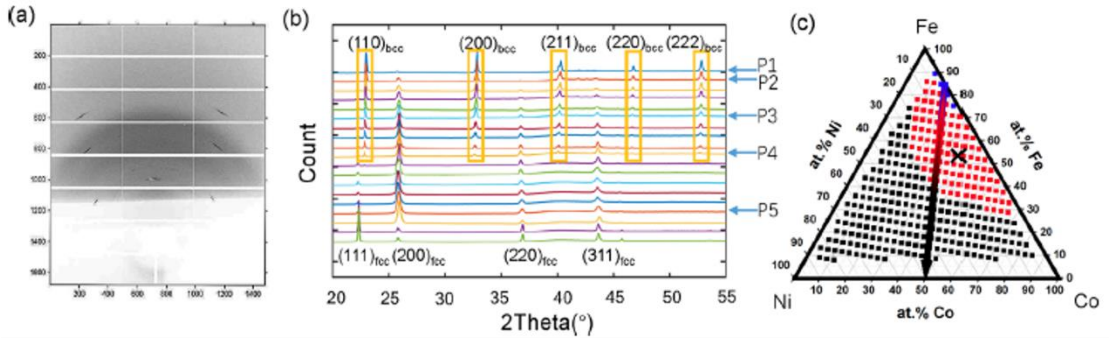
各点化学成分的确定是有 X 射线荧光光谱确定的，因为 Fe、Co、Ni 三种元素的相对原子质量较近，三种元素的空间分布较为线性化，此现象与实验设计的初衷一致。

两相区中的一个能谱的面心立方和体心立方的晶格常数预测值，分别与 Ni 的面心立方 PDF 卡片（no.04-0850）的值 0.3523nm 及 Fe 的体心立方 PDF 卡片（no.06-0696）的值 0.2866nm 比较后，取为 0.356 和 0.285nm。早期研究表明，Fe 和 Co 混入面心立方的 Ni 会导致晶格常数的增加，体心立方的 Fe 的晶格常数随着 Co 的原子百分比增加至 25%一直呈正相关的关系，之后随着 Co 的原子百分比的增加而递减。因此，合金化后的晶格常数的畸变与合金化的定性趋势是一致的。

为了自动确定相区域，采用层次聚类去给衍射谱分组。Cosine 距离被用来计算两条衍射谱之间的相似度，用 D 表示衍射谱 Pm 和 Pn 之间的距离，计算公式如下：

$$D_{cosine}(P_m, P_n) = 1 - \frac{\sum_{i=1}^{n}(p_{m_i} \cdot P_{n_i})}{\sqrt{\sum_{i=1}^{n} P_{m_i}^2} \cdot \sqrt{\sum_{i=1}^{n} P_{n_i}^2}}$$

P_{m_i}和P_{n_i}代表矢量P_m和P_n的第 i 个元素。



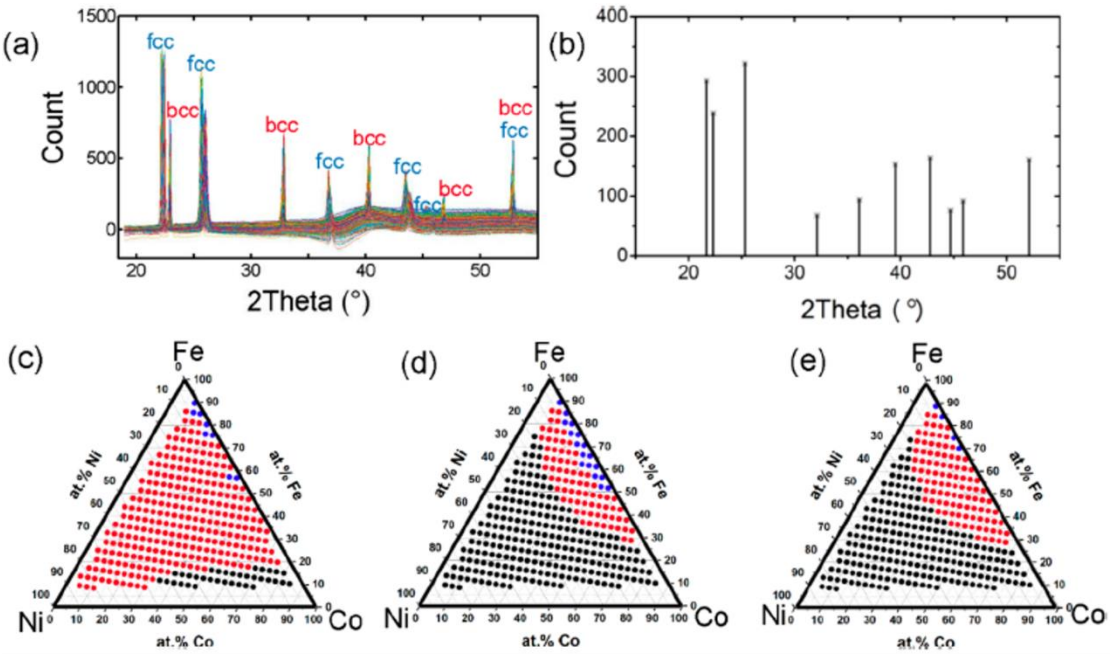
（a）曝光时间为1s的典型衍射照片

（b）延扫描线并经转换后的XRD能谱

（c）用600℃下热处理的材料芯片手动构建的成分相图（蓝点，bcc；红点，bcc+fcc；黑点，fcc）

图二

图 3a 是 600℃下热处理的材料芯片所有 XRD 衍射谱的集合。在聚类的过程中采用了三种策略。首先比较能谱每个衍射角的强度（全局策略），据此在成分空间识别出三个类别。尽管聚类结果并不违反吉布斯相规则，但是两相区的尺寸远大于图 2c 手绘的结果。如此大的误差主要是由于整个能谱不规则的背底造成的，导致聚类分析中引入较多的噪声。另外，由于相分辨造成的峰的偏移和峰高的改变也被认为是错误聚类的一个主要因素。



（a）fcc和bcc典型的XRD衍射能谱

（b）从表a识别出的峰位和强度

（c）使用全谱策略的聚类结果

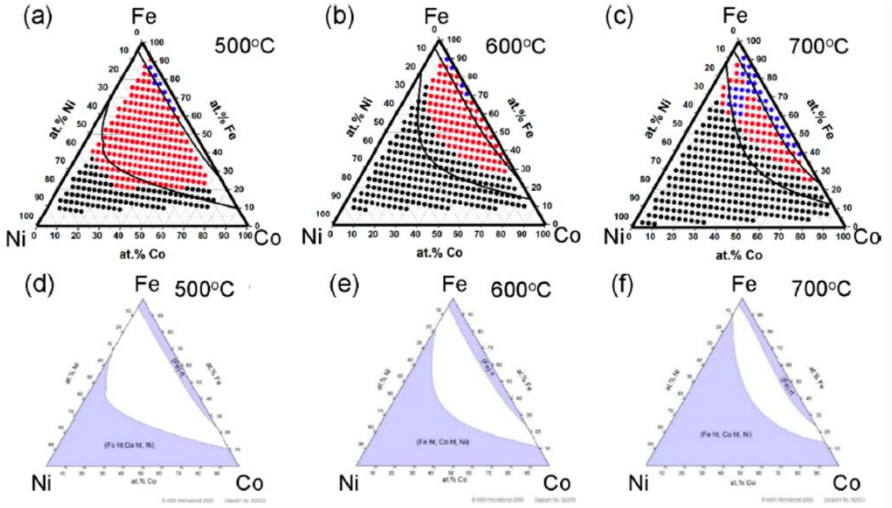
（d）使用部分能谱的聚类结果

（e）使用部分能谱策略进一步优化后聚类结果

图四

其次，抛开全谱策略，在遵循标准预处理流程的条件下，峰的位置和强度被提取出来。从每个能谱中提取出一组离散的峰位和强度。依赖于成分，芯片上不同位置得到的相同衍射峰可能位于标准衍射峰附近的一定范围内（峰位的偏移）。峰位偏移的范围通过取所有能谱中相同峰位的平均值来确定的。所有的能谱均被转换为布尔类型的变量表，该表包含所有能谱中所有标准的峰位。相似度 D 是基于布尔向量 P 计算的。使用部分能谱的聚类结果相比于使用全能谱聚类的结果更加真实。相比于手动分类的 317 条能谱中有 282 条被正确分类

（89%），这是较全谱聚类结果的一大进步。



（a）、（b）、（c）分别是500、600、700℃下热处理后的材料芯片构建的成分相图（蓝点，bcc；红点，bcc+fcc；黑点，fcc）。（d-f）分别是ASM相图数据库对应温度下的相图

图五

为了进一步提升正确率，采用了一套自动化的相标记。已知每个峰属于那种晶体结构，能谱可以被标记为 fcc、bcc 和 fcc+bcc。通过这种方式，在相图边界容易分类的错误的点也能被识别出来。基于部分能谱的聚类结果和相标签的总和结果，317 条能谱中有 310 条能谱被正确分类，这在正确率上又是一大进步。

相同的不做流程被应用于构建 500 和 700℃下的成分相图。与手动分类的结果相比，306 条能谱中正确分类 298 条（97.3%），336 条能谱中正确分类 312 条（92.9%）。聚类结果与 ASM 相图数据库中的相图结果高度一致。当前红点区域的两相区可能会随着退火温度的下降而扩张。700℃下热处理的成分相图中红点区域里的少数离散蓝点是由于 fcc 较弱的特征峰造成的。这些越界的点可以通过认为的督察排除掉。

本文使用 Fe-Co-Ni 三元合金系统来阐明快速构建成分相图的方式。100nm 厚、等温热处理的材料芯片薄膜在同步光源的微束 XRD 和 XRF 进行表征。每个样本点的曝光时间为 1s。基于层次聚类的两步处理能构建平均正确率超过 95%的成分相图且于 ASM 相图数据库结果一致，从而验证了方法的有效性。期待这种方法也能成功的应用到其他材料系统中。