## 多目标进化聚类和此算法在患者分层上的应用

**摘要：**

患者分层在某些医疗上很有用。现有的一些算法在对处理这个问题上有一些局限性，比如：实验噪音，高维度和不可解释。现有的算法通常仅使用单目标来确定聚类质量。这样种单目标的函数不是很适合对于所有的数据集。

在本文中，我们要使用一种新的框架叫做通过快速寻找密度峰的多目标聚类算法（multiobjective clustering algorithm by fastsearch and find of density peaks）在这种框架中，候选的种群在多目标下得到演化选择特征，比情怯自动评估聚类密度（In the proposed framework, a parameter candidate population is evolved under multiple objectives to select features and evaluate clustering densities automatically.）。

五个聚类的评估标准是：紧实性（compactness），分离性（separation）Calinski–Harabaszindex,和 Davies–Bouldin index,和 Dunn index。这五个评估标准领导了多特征的进化聚类。

基于分解的多目标聚类算法被用来同时优化五个目标函数。然后我们也和一些别的算法做了对比，事实证明，我们所提出的算法在结果和时间复杂度上都比别的算法好很多。

**介绍：**

**（第一段）**

患者分层是一个关键的改变对于现代医学在基因药物上面的发展。在过去几年里，新的基因测序技术已经使得我们能够大量的测出不同的分子集合。第一段就是讲的这个患者分层对于现代医学的意义

（第二段）

下面是一些过去的，关于患者分层的研究，有7个标出来了，先不看了

以上的算法都将不同的特征平等对待，这就导致了他们在高维度分子数据上表现的不是很好。因此需要

1. 将不同的特征分配一个不同的权重
2. 使用多目标的聚类

（第三段）

为了量化不同的特征的影响，距离权重经常被合并到特征中，很多传统的聚类算法已经结合了权重。比如有8个作者已经做过相关的工作。对他们的工作表示肯定，我们也要设计一个合适的距离权重用来减少实验噪音和冗余的数据，做出一个更反映真实的模型。

（第四段）

过去的很多论文都是单目标进化聚类算法，比如一堆

（第五段）

最近一种新的算法被提了出来：(CDPs) 。聚类中心对与别的点密度更大，聚类中心距离别的距离中心的距离比别的非聚类中心距离别的聚类中心的距离大。这个算法已经被成功的用于分子数据聚类。但是这个算法在处理高维度的数据时候表现的不是太好。

CDP算法在一开始就计算了欧式距离，随着维度的增加，虽然样本点很不一样，但是他们所表现在欧式距离上没有太大的区别。这将会导致计算*ρi 和 δi* 的时候很难进行下去。

同时截断距离Dc的选择也应该是一个截断距离是一个启发式的距离，不应该固定他的大小，这个距离一边选择是：让所有的Population中的1%~2%的点落在圈里面

基于以上的分析我们设计了一个新的算法叫做multiobjective CDPs (MOCDPs) for clustering patient stratification data. 这个算法通过结合权重和启发式的截断距离解决了CDPs的缺点。选取五个聚类有效性指标作为内在聚类的目标函数。

最后，给出了基于分解的多目标DE算法

(MOEA/D-DE)同时优化这五个目标函数，以稳健的方式将患者分层为亚型。

**方法：**

1. CDPs 介绍（这是一个聚类方法）

一个详细的博客：

<https://blog.csdn.net/matrix_space/article/details/82748218>

CDPs[29]是一种根据元素相似性对其进行分类的新算法，该算法已于2014年成功应用并发表在《科学》杂志上。CDP基于两个假设

1. 集群中心是一个高度密集的数据区域
2. 集群中心与其他集群中心的距离相对较远。

每个数据点的ρi密度和距离δi计算公式如下：

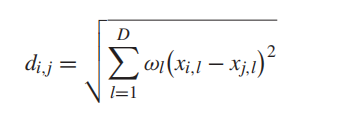
1. 改进

在CDP算法中有两个重要的输入参数

1. 相异度矩阵
2. 截断距离

对于相异度矩阵，CDP使用欧氏距离测量特征成分，这种方法的一个显著的问题就是选择特征。CDP算法不能自动的选择特征因为他将所有的特征都平等对待在评估相异程度的时候。但是这是不现实的将所有的特征都平等对待。事实表明，有效的聚类算法都是用了特征的子集。除此之外，其中还有一些特征是多余的，甚至是不利于聚类的。

在这个论文中，我们首次提出增加一个权重相异度矩阵到CDP中计算距离



*di,j* 是xi和xj的距离

*ω* = {*ω*1*, ω*2*,...,ωl*} 是权重向量，权重向量的和是1

因为我们不能事先知道这个权重是多少，所以我们需要使用多目标优化的方法来优化这个权重。同时，我们使用一种启发式的方式来选择*dc* 让落在圈里面的样本占据所有样本的1%-2%左右。

其他的不同于CDP的，原来的CDP使用一种评估函数，现在我们使用所中评估函数来同时优化我们的算法。

1. 快速寻找和密度峰多目标聚类算法

这一段就是介绍切比雪夫方法的

在这部分，我们提出基于分解位置的MOCDP以解决患者背景下的聚类问题分层。 MOCDP利用基于分解的框架，该框架类似于多目标进化框架基于分解的算法（MOEA / D）[32]。 这种算法开始于具有相似权重和患者分层数据的截止距离。之后，它分解了提出的多目标聚类问题使用tchebychev方法分解成一些标量子问题，具有用于目标函数的一组均匀分布的权重向量λ。 之后，每个子问题都会通过MOEA / D-DE仅使用来自其邻居的信息子问题。 每个子问题都有其最佳解决方案，因此人口众多。 然后，比较新个体原始解具有相同的权重向量λ和它的相邻子问题。

1. 多目标优化

介绍了多个f（x），pareto优化，pareto最优解集

1. 目标函数

这个论文使用了五个目标函数

五个聚类的评估标准是：紧实性（compactness），分离性（separation）Calinski–Harabaszindex,和 Davies–Bouldin index,和 Dunn index。这五个评估标准领导了多特征的进化聚类。

1. 多目标聚类算法

我们结合了CDP和多目标进化算法生成一个新的算法

切比雪夫方法就是用来分解多目标问题为单目标问题的一种方法。