**研 究 生 课 程 论 文**

**(2018-2019学年第一学期)**

**粒子群K-均值算法**

**研究生：曹业伟**

**提交日期： 2019年6月28日 研究生签名：**

|  |  |  |  |
| --- | --- | --- | --- |
| **学 号** | **201820137791** | **学 院** | **软件学院** |
| **课程编号** | **S0001073** | **课程名称** | **高性能算法设计与分析** |
| **学位类别** | **硕士** | **任课教师** | **黄翰 教授** |
| **教师评语：** | | | |
| **成绩评定： 分 任课教师签名： 年 月 日** | | | |

[[1]](#footnote-1)

粒子群K-均值算法

曹业伟

*摘要*— 生物信息学中产生了大量的高维数据集，利用聚类分析这种探索式的分析方法，可以较好的找到生物数据的之间内在联系，利用启发式的粒子群算法结合K-均值算法，寻找出好的初始质心位置，减少K-均值的迭代次数，对高维数据集有不错的效果。也分析了粒子群K-均值算法相关不足之处，为下一步改善算法提出了思路。

*关键词*—聚类分析 生物信息 粒子群 K-均值

# 引言

随着生命科学前所未有的高速发展，产生的生物信息也急剧增加，为了处理这些庞大的数据集，人们巧妙的将计算机科学等分析工具和方法引入到生命科学研究中来，形成了生物信息学[1]

癌症是一种通过多种途径演变的异质性疾病，涉及多种癌基因和肿瘤抑制基因活性的变化。这些变化的基础是产生复杂分子和细胞现象类型的大量和多样的体细胞替代物，影响每个个体肿瘤的行为和反应治疗。由于突变和分子机制的多样性，结果差异很大。因此，将机器学习的方法对癌症病人的基因表达情况进行研究，对于识别病症亚型，癌症基因表达的内在联系识别有着重要的意义。[2]

对于基因数据常常是高维数据这一特点，国内外有学者采用了各种不同降维算法，相似度学习方法，来降低待测样本的维度例如相似度网络融合SNF[3]，icluster+等，也有学者根据多核学习理论提出了一种优化求解相似度矩阵的方法[4]。

以上这些方法，对于高维数据集都有着不错效果，但是优化过程比较复杂，涉及到很多数学理论知识。因此，我们想要采用启发式算法，同时结合较为简单降维算法，对高维数据集进行聚类分析。

# 理论方法

## K-均值算法

在所有的聚类算法之中，K-means聚类算法可以说是众所周知的简单实用的经典聚类算法[5]，数学定义如下：

给定样本集 ,K-means算法的目的就是对于根据算法划分个簇有最小化的平方误差。



其中的 是每个簇自身的质心，也就是这整个簇的均值向量。[11] 周志华. 机器学习[M]. 北京：清华大学出版社，2016：202-220.

算法的大体步骤如下：首先，它根据某种算法先选取个样本，把这个样本作为这个簇的样本均值向量，也就是这一个簇的质心。对于其余每个样本，计算出样本与簇质心之间的距离，将这个样本分配给最相近的簇。在这之后下一步计算每个簇的新的心。一直不断重复这个过程，直到这个过程迭代到目标函数收敛[6]。

算法伪代码如下：

|  |
| --- |
| 1:从样本随机选择样本作为初始质心  2:REPEAT  3: 令  4: FOR  DO  5: 计算出样本与每个质心的距离  6: 根据距离最近质心决定分入的簇的标记  7: 把样本分入相应的簇  8: END FOR  9: FOR  DO  10: 计算新的质心向量  11: IF  THEN  12: 把当前的质心向量更新为  13: ELSE  14: 当前的质心向量不变  15: END IF  16: END FOR  17:UTIL达到当前输出条件 |

对于值的选取方法，目前多靠人工根据数据情况进行选取，也有一种比较可行的方法，是Elbow Method：这种方法是要绘制K-均值的代价函数和聚类簇数目的关系图线，这样就可以选取这个曲线拐点的值作为最好的聚类簇数目。但是这种方法在实际应用之中出现的次数比较少，因此我们不在这里过多的介绍。[7]根据经验规律认为最佳的聚类数应该在２与槡Ｎ之间，其中Ｎ为数据空间中的所有数据点的个数。对于我们来说，还是提倡从实际问题入手，人工经过一定的试验，通过多次试验来决定聚类中心。对于初始质心有学者提出过多次初始化质心坐标并聚类[8]，也有学者提出基于密度的解决方法。[9]

## 粒子群算法

粒子群优化算法(PSO)[10]是一种进化计算技术，它模拟鸟群的运动行为建模

假设这里有个粒子，所以种群即为，每一个粒子的位置为，速度是，每一个粒子的个体最优值是，全局最优值是，速度和位置的更新公式如下：





公式中代表迭代次数，是惯性权重，是学习率，是两个独立的随机参数，取值在0到1之间，算法的整体流程如下图所示，首先随机初始化种群，然后计算每个粒子的适应度，得到个体最优值和全局最优值，接着按照公式更新每个粒子的位置和速度，最后看能否达到迭代终止条件，若不能则继续从第二步开始迭代。

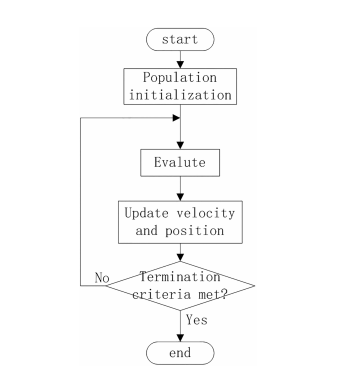


图1 粒子群迭代过程

## 粒子群K-均值算法

粒子群算法是一种全局搜索算法，通过适当调节其参数，加快收敛速度，提高全局寻找最优解的能力。但是，粒子群算法在接近最优解时收敛速度变得过慢。[11]相反，K-均值算法具有快速收敛于局部最优解的能力，但它寻找全局最优解的能力较弱，容易收敛到局部最优[12]。我们把粒子群和K-均值算法整合起来，前期先使用粒子群算法寻找全局搜索整个空间内的大致的最优解，之后换成K-均值算法进行进一步计算具体的质心坐标。[13]

我们在进行优化过程中，把一个粒子视为一组质心坐标。选取一定数量的粒子，进行随机初始化。样本集，维数为假设这里有个粒子，所以种群即为，每个粒子的位置为，为一个维的向量。

采用的评价标准是Calinski-Harabasz分数。[14]聚类效果的目标函数如下



其中的是训练集样本数，是类别数。是类别之间协方差矩阵，是类别内部数据的协方差矩阵。为矩阵的迹。得分越高，聚类效果越好。

算法步骤如下：

1. 初始化粒子群，输入相关的参数值例如最大迭代次数，粒子数目，聚类数目等
2. 随机从所有数据中选择个作为质心，并把这个粒子作为一个粒子，重复生成个粒子
3. 计算每个粒子适应度函数值，即CH得分，得分越高聚类效果越好，进而不断更新和
4. 根据公式计算出每个粒子的速度，并且更新粒子位置，不断迭代，直到达到终止条件。

# 实验

## 数据来源

实验数据集是来自TCGA的低级胶质瘤数据集一共282个病例样本，相关基因有20890个，每个基因有基因表达量，拷贝数替代，甲基化，点变异四种测量值。

对相关的四个数据集进行清洗融合，大于10的值更改为10，小于-10的值更改为-10，之后归一化到0-1范围内，缺失值填充为0.5。采取PCA[15]降维的方式将四个特征进行融合，最后得到一个的样本矩阵

## 实验过程

对于实验数据集，根据原论文测试结果，聚类数目3时聚类效果最好，我们选取聚类数目为3，采用了两种思路去进行聚类。一种是直接在原有的样本矩阵上聚类；另一种是我们在进行PCA和t-SNE[16]降维之后，对二维数据进行聚类分析。

对于原始数据直接进行聚类，粒子数为50，最大迭代次数为100。经过计算得到的CH分数为49.97。

采用PCA和t-SNE分两步对原始数据集降维，首先利用PCA将原始数据集降维到80维，再利用t-SNE降维到2维并进行可视化，t-SNE的困惑度为10，学习率为150，最后进行粒子群K-均值聚类，粒子数为50，最大迭代次数为100。可以比较明显的区分为3类数据。

聚类结果如下图所示：

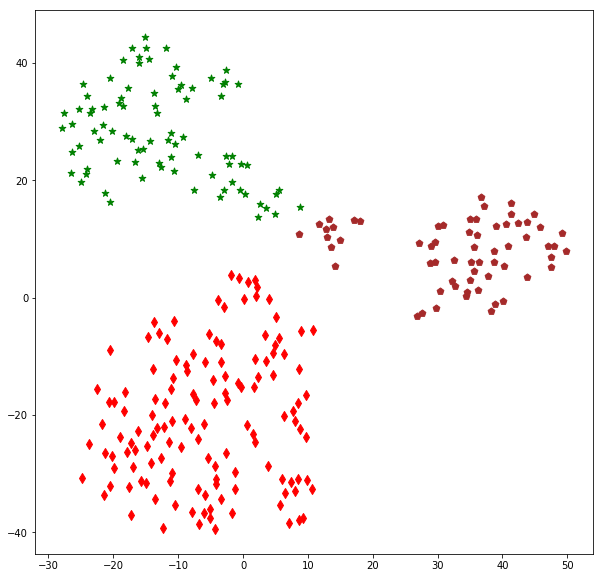


图2 粒子群K-均值算法

其次，对比了我们寻找到的最优质心和最后聚类完成的质心，比较这两种的质心聚类的效果，如下表所示：

|  |  |  |
| --- | --- | --- |
|  | 原始数据 | 二维数据 |
| 初始质心CH得分 | 37.45 | 262.63 |
| 最终质心CH得分 | 49.97 | 587.43 |

表1 原始数据与二维数据得分对比

通过对比可以发现，对于高维度的数据，粒子群K-均值算法找到的质心位置和最后聚类找到的质心位置比较接近，聚类效果也相差不多，减少了K-均值的计算次数，能提供比较好的初始质心值，对于二维数据，也有不错的效果。

最后我们对比和原论文CIMLR的聚类效果

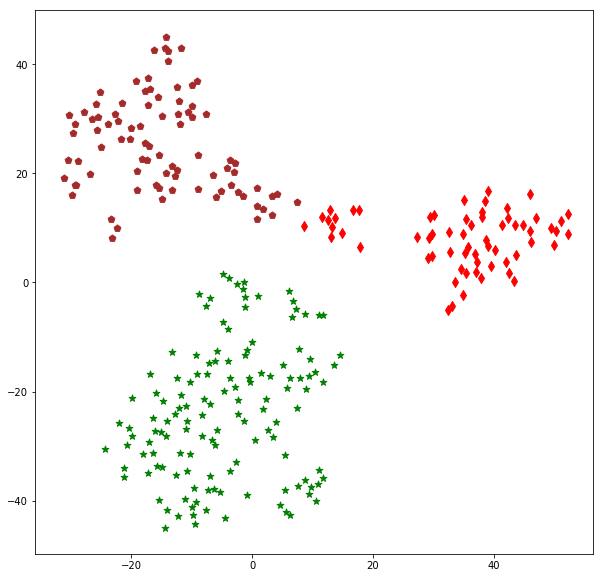
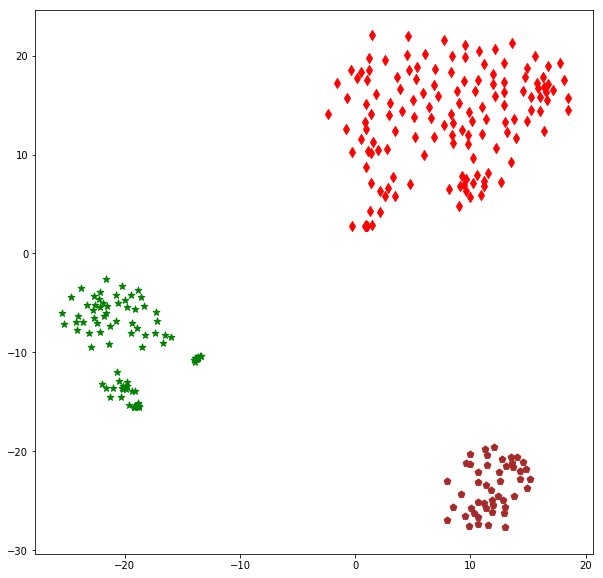


图3 左图为CIMLR，右图为粒子群K-均值算法

CIMLR的聚类CH得分1356.70，聚类效果更好，但是算法较为复杂，依赖通过多核学习形成的相似度矩阵进行降维聚类，步骤较多，难以应用于高维数据。粒子群K-均值算法算法适应程度比较广，从高维数据到低维数据都可以使用，应用于高维数据效果要更好一些。

## 实验结论

粒子群K-均值算法，适应于比较广的数据集，从高维到低维都可以采用，并且可以减少高维数据集K-均值算法迭代次数，更好地找到最终质心的位置。相较于CIMLR，算法更加简单明了，迭代求解的步骤也更少，并且聚类的区分度也不错。

但同时粒子群K-均值算法也暴露出了一些问题，在粒子群求解质心的过程之中，比较容易早熟现象产生，得到的初始质心比较差。在降维过程中，也比较依赖降维的手段，单一的PCA和t-SNE降维的效果都不算好，只有联合起来降维，进一步聚类的效果才比较好。同时对于四种特征的融合只用了PCA，没有考虑到实际生物学的背景意义。对于样本与样本之间生成的的相似度矩阵，降维前后粒子群K-均值算法均没有很好的聚类效果，也是本次实验不足之处。

# 总结

当前在生物信息领域，对基因数据的分析是一个热点。由于基因数据数量庞大而且复杂，而聚类分析以“物以类聚”基本思想，基于探索式的数据分析方法，尤其当数据量较大时，可以取得很好的分析结果，所以成为处理生物信息数据时最常用到的一种有效手段

生物信息的聚类主要目的是利用探索式的数据分析方法，找出疾病信息的共性和内在联系。粒子群算法具有简单性和和自组织性，适应性较强，算法的实现也比较简单，K-均值算法也是一种优秀的聚类算法，思想简单，聚类效果较好，本文将两者结合，利用粒子群定位出K-均值算法的初始质心，减少K-均值的迭代次数，同时也能在高维的数据集有不错的效果。但粒子群K-均值算法也有诸多不足，粒子群早熟的特点容易产生偏离比较大的质心，对于相似度矩阵的应用效果不好，参数的选择缺乏通用性等，这也是以后要进一步改善的问题。

参考文献

1. 靳艳虹. 基于 PSO 的基因表达数据聚类研究. MS thesis. 中南大学, 2013.
2. Ramazzotti, Daniele, et al. "Multi-omic tumor data reveal diversity of molecular mechanisms that correlate with survival." *Nature communications* 9.1: 4453.*.* 2018
3. Wang, Bo, et al. "Similarity network fusion for aggregating data types on a genomic scale." *Nature methods* 11.3: 333 2014.
4. Wang, Bo, et al. "Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning." *Nature methods* 14.4: 414 2017.
5. 吴夙慧, et al. "K-means 算法研究综述." *数据分析与知识发现* 27.5: 28-35 2011.
6. 曲建华, 邵增珍. "多种群协同进化的 K-means 聚类算法." *南京师大学报 (自然科学版)* 33.3: 122-126 2010
7. Rezaee, M. Ramze, Boudewijn PF Lelieveldt, and Johan HC Reiber. "A new cluster validity index for the fuzzy c-mean." *Pattern recognition letters* 19.3-4: 237-246 1998.
8. Hart, Peter E. Pattern classification and scene analysis. 1973.
9. Katsavounidis, Ioannis, C-C. Jay Kuo, and Zhen Zhang. "A new initialization technique for generalized Lloyd iteration." *IEEE Signal processing letters* 1.10: 144-146 1994.
10. ]Eberhart, Russell, and James Kennedy. "Particle swarm optimization." *Proceedings of the IEEE international conference on neural networks*. Vol. 4. 1995.
11. 郑晓鸣, 吕士颖, 王晓东. "基于免疫粒子群优化的聚类算法." *计算机工程* 34.15: 179-181 2008.
12. ]Selim, Shokri Z., and Mohamed A. Ismail. "K-means-type algorithms: A generalized convergence theorem and characterization of local optimality." *IEEE Transactions on pattern analysis and machine intelligence* 1 : 81-87 1984.
13. 徐辉, 李石君. "一种整合粒子群优化和 K-均值的数据聚类算法." *山西大学学报: 自然科学版* 4 : 518-523 2011.
14. ]Caliński, Tadeusz, and Jerzy Harabasz. "A dendrite method for cluster analysis." *Communications in Statistics-theory and Methods* 3.1 : 1-27 1974.
15. Jolliffe, I. T. . "Pincipal Component Analysis." *Journal of Marketing Research* 25.4:513 2002.
16. Maaten, Laurens van der, and Geoffrey Hinton. "Visualizing data using t-SNE." *Journal of machine learning research* 9.Nov: 2579-2605 2008.

1. [↑](#footnote-ref-1)