高维聚类算法开发文档

# 1 算法原理

## 1.1 算法概述

将物理或抽象对象的集合分组称为由类似的对象组成的多个类的过程被称为聚类。高维聚类分析与传统聚类分析的最主要差别就是高维度。高维数据聚类是聚类技术的难点和重点。高维数据聚类已成为数据挖掘中的一个重要研究方向。因为随着技术的进步使得数据收集变得越来越容易，导致数据库规模越来越大、复杂性越来越高，如各种类型的贸易交易数据、Web 文档、基因表达数据等，它们的维度（属性）通常可以达到成百上千维，甚至更高。但是，受“维度效应”的影响，许多在低维数据空间表现良好的聚类方法运用在高维空间上往往无法获得好的聚类效果

传统的聚类算法可分以下五类 ：①划分方法②层次方法③基于密度的方法④基于网格的方法⑤基于模型的方法。它们已经比较成功的解决了低维数据的聚类问题。但是由于实际应用中数据的复杂性，在处理许多问题时，现有的算法经常失效，特别是对于高维数据和大型数据的情况。因为传统聚类方法在高维数据集中进行聚类时，主要遇到两个问题。①高维数据集中存在大量无关的属性使得在所有维中存在簇的可能性几乎为零；②高维空间中数据较低维空间中数据分布要稀疏，其中数据间距离几乎相等是普遍现象，而传统聚类方法是基于距离进行聚类的，因此在高维空间中无法基于距离来构建簇。

目前一般使用两种方法解决以上问题：**（1）特征转换，（2）特征选择 /子空间聚类。**

**特征转换**是一种传统的方法，包括**主成份分析**和奇异值分解等策略。该方法通过线性合并将原数据集的维合并至k个新维，使得诸如k～均值一类的传统算法能在这k个新维中进行有效聚类，从而达到减少维的目的。

**特征选择 /子空间聚类**和特征转换不同，它只在那些相关的子空间上执行挖掘任务，因此它比特征转换更有效地减少维。特征选择一般使用**贪心策略**等搜索方法搜索不同的特征子空间，然后使用一些标准来评价这些子空间，从而找到所需的簇。

子空间聚类算法拓展了特征选择的任务，尝试在相同数据集的不同子空间上发现聚类。和特征选择一样，子空间聚类需要使用一种搜索策略和评测标准来筛选出需要聚类的簇，不过考虑到不同簇存在于不同的子空间，需要对评测标准做一些限制。选择的搜索策略对聚类结果有很大的影响。根据搜索的方向的不同，可以将子空间聚类方法分成两大类：**自顶向下的搜索策略和自底向上的搜索策略。**子空间聚类是实现高维数据集聚类的有效途径，它是在高维数据空间中对传统聚类算法的一种扩展，其思想是将搜索局部化在相关维中进行。

## 1.2 算法组成与步骤

算法步骤：

**基于特征转换法（主成分分析）以及k-means算法实现高维聚类算法流程：**

1）对原始数据进行标准化处理。设xij表示第i(i= 1,2,… ,n)个被测量第j(j= 1,2,… ,m)个指标的指标值，则xij的标准化值为：,其中，。是经过标准化变换后得到的数据矩阵：。

2)求标准化数据的相关矩阵，。

3)计算相关矩阵的特征值λ1≥λ2≥ … ≥λm(满足)和其对应的特征向量u1,u2,… ,um。

4)确定主成分个数。找出前p个主成分,满足。

5)前p个主成分对应的特征值组成的矩阵为Um×p,则缩减后p个综合指标为。

6）由主成分分析求得的综合指标矩阵作为k-means算法的输入是样本集D={x1,x2,...xn}，同时主成分个数p为k-means的数据集维数，设定：聚类的簇数k,最大迭代次数N，新老质心变动距离的阈值Q

输出是簇的划分C={C1,C2,...Ck}

7）从数据集D中随机选择k个样本作为初始的k个质心向量： {μ1,μ2,...,μk}

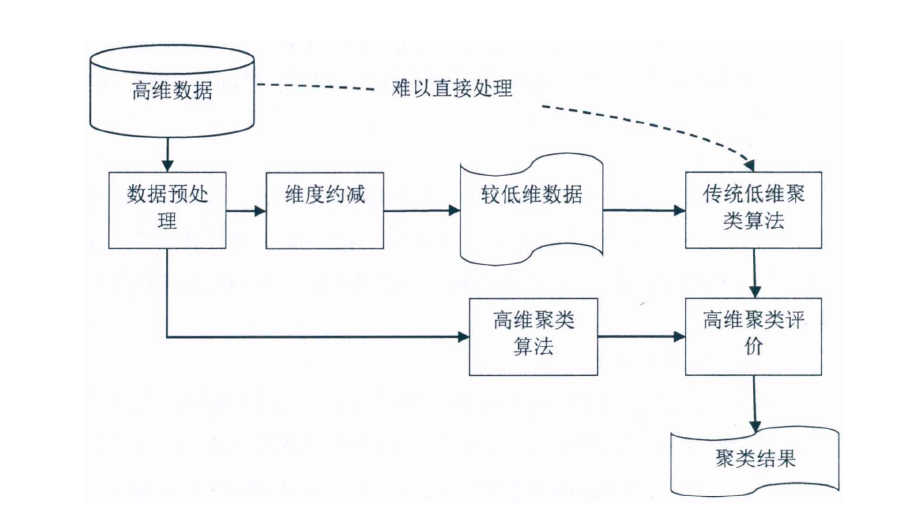
8）对于n=1,2,...,N：将簇划分C初始化为Ct=∅t=1,2...k

9）对于i=1,2...m,计算样本xi和各个质心向量μj(j=1,2,...k)的距离：dij=||xi−μj||，将xi标记最小的为dij所对应的类别λi。此时更新Cλi=Cλi∪{xi}

10）对于j=1,2,...,k,对Cj中所有的样本点重新计算新的质心μj=1|Cj|∑x∈Cjx

11）如果所有的k个质心间的新老距离小于输入的阈值或者达到最大迭代次数，则转到步骤12。否则重复执行步骤9-11）

12）输出簇的划分结果C={C1,C2,...Ck}



## 1.3 算法核心参考文献

基于超图划分的高维数据聚类方法研究\_王田

基于稀疏表示的混合属性数据聚类关键技术研究\_邵晓晨

# 2 算法开发

## 2.1算法封装（输入与输出）

|  |  |  |
| --- | --- | --- |
|  | 名称 | 属性 |
| 输入：样本集 | Date | double[][] |
| 输入：PCA阈值 | threshold | double |
| 输入：聚类的簇数k | k | int |
| 输入：最大迭代次数N | MAX\_ATTEMPTS | int |
| 输入：新老质心变动距离的阈值Q | MIN\_CRITERIA | double |
| 输出：簇的划分结果 | res | double[] |

## 2.2算法核心函数及说明

组成关系图参考





## 2.3 算法组成与执行流程

算法由HdimCluser、Pca、Kmeans、KmeansDate、KmeansParam、KmeansResult六个类组成，其中HdimCluser为高维聚类算法类、Pca 为主成分分析类、Kmeans类为Kmeans算法的具体实现、KmeansDate为数据类、KmeansParam为参数结构类、KmeansResult为测试结果类。

执行流程：

主成分分析：

IO流输入初始数据

数据标准化

样本相关系数矩阵计算

样本相关系数矩阵的特征值计算

样本相关系数矩阵的特征向量计算

获取主成分

求取缩减后的综合指标矩阵

Kmeans:

将主成分分析得到的结果作为Kmeans数据集输入，同时数据维度等于主成分个数。

初始化参数结构、设置聚类中心点的初始化模式。

设置要分的簇的个数k，做kmeans计算。

计算最终聚类结果

# 3 算法验证

## 3.1 验证算例说明

参考高维聚类算法文件夹的test2被测数据集以及与其对应的result2数据集聚类结果

## 3.2 验证结果说明

输入的测试数据为：

80.4000 96.3000 76.9000 88.7000 11.1000

90.0000 96.1000 75.8000 90.3000 10.6000

93.8000 97.2000 97.8000 88.4000 10.1000

88.4000 96.2000 91.2000 90.9000 12.5000

98.4000 96.5000 86.7000 92.7000 12.2000

77.7000 83.9000 75.2000 84.3000 7.6000

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

原始数据标准化矩阵为：

-0.9804 0.3759 -0.7458 -0.1799 0.2357

0.2393 0.3371 -0.8625 0.3772 -0.0471

0.7221 0.5510 1.4705 -0.2843 -0.3299

0.0360 0.3565 0.7706 0.5861 1.0276

1.3066 0.4148 0.2934 1.2128 0.8579

-1.3235 -2.0353 -0.9261 -1.7118 -1.7440

数据标准化完毕

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

计算样本相关系数矩阵：

1.0000 0.6717 0.6447 0.7865 0.5971

0.6717 1.0000 0.5041 0.8198 0.8325

0.6447 0.5041 1.0000 0.3567 0.4077

0.7865 0.8198 0.3567 1.0000 0.9323

0.5971 0.8325 0.4077 0.9323 1.0000

样本相关系数矩阵计算完毕

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

计算相关系数矩阵的特征值：

0.0005 0.0000 0.0000 0.0000 0.0000

0.0000 0.1790 0.0000 0.0000 0.0000

0.0000 0.0000 0.3433 0.0000 0.0000

0.0000 0.0000 0.0000 0.8112 0.0000

0.0000 0.0000 0.0000 0.0000 3.6660

样本相关系数矩阵的特征值计算完毕

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

计算相关系数矩阵的特征向量：

-0.3890 0.1477 -0.7354 -0.2889 0.4501

0.0090 0.7786 0.3850 0.1443 0.4740

0.2060 -0.2381 0.3664 -0.8086 0.3358

0.7225 -0.2377 -0.2636 0.3372 0.4881

-0.5331 -0.5088 0.3274 0.3579 0.4709

样本相关系数矩阵的特征向量计算完毕

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

样本相关系数矩阵的特征值矩阵(一维)为：

0.0005 0.1790 0.3433 0.8112 3.6660

将样本相关系数矩阵的特征值矩阵中的数据从大到小排序:

3.6660 0.8112 0.3433 0.1790 0.0005

选取前N个主成分对应的特征值数组的索引值为：

[2, 3, 4]

获取主成分完毕

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

前N个主成分对应的特征值的特征向量矩阵为：

-0.7354 -0.2889 0.4501

0.3850 0.1443 0.4740

0.3664 -0.8086 0.3358

-0.2636 0.3372 0.4881

0.3274 0.3579 0.4709

缩减后的综合指标矩阵为：

0.7171 0.9643 -0.4903

-0.4771 0.7873 0.1398

0.1868 -1.5321 0.7858

0.5750 -0.0167 1.2139

-0.7325 0.1612 1.8791

-0.2693 -0.3639 -3.5282

求取缩减后的综合指标矩阵完毕

主成分分析计算完毕

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

初始质心坐标为：

-0.4771 0.7873 0.1398

0.1868 -1.5321 0.7858

-0.7325 0.1612 1.8791

第1次迭代质心坐标为：

-0.0098 0.4625 -1.2929

0.1868 -1.5321 0.7858

-0.0788 0.0722 1.5465

第2次迭代的质心坐标为：

-0.0098 0.4625 -1.2929

0.1868 -1.5321 0.7858

-0.0788 0.0722 1.5465

迭代完毕

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

聚类个数=3

迭代次数=2

退出迭代时的新旧质点间最大距离（小于阈值）=0.0

每个聚类包含的点的个数:

3 1 2

原始数据对应的分类为:

属于0簇的数据：

0.7171 0.9643 -0.4903

-0.4771 0.7873 0.1398

-0.2693 -0.3639 -3.5282

第0簇数据归纳完毕

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

属于1簇的数据：

0.1868 -1.5321 0.7858

第1簇数据归纳完毕

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

属于2簇的数据：

0.5750 -0.0167 1.2139

-0.7325 0.1612 1.8791

第2簇数据归纳完毕

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

计算完毕

流程图参考

