# 附录Ⅴ.使用kmeans算法进行聚类分析

## 背景介绍

我们使用SVDFeature分析了数据之后，得到了每个作品的特征向量，由此我们希望能通过这些特征向量来对作品使用聚类分析进行分类，希望在之后的评分或者对评委的指南中起到作用。

我们使用kmeans算法来进行聚类分析，其中两个向量之间的距离我们取余弦相似度，因为作品之间的类别即可以抽象为向量方向的类别

## Kmeans算法

### 基本描述

kmeans算法主要为以下步骤

1.选取k个点为初识质心

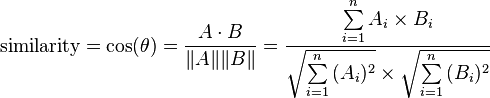
2.将每个点选取最合适的质心，形成k个簇

3.重新计算质心

4判断簇的变化，选择继续2或停止。

### 距离函数和质心计算

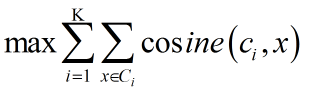
距离函数为余弦相似度：



质心计算即为簇中所有向量的平均

### 算法停止条件

目标函数达到最优解或者达到最大的迭代次数，我们用到是余弦相似度所以目标函数为：



### k的选取

我们分别尝试不同的k，然后观察目标函数的变化，通常情况下，在k的数量小于真实值时，目标函数会迅速变化而当超过真实值后，目标函数会平缓变化。

经过对数据的分析，我们选取最佳的k为8。

## 分类结果

具体分类结果可见cu2012.txt已经cu2013.txt

为了更清晰的反应分类效果，我分别绘制了2012和2013年最终分类结果每个簇质心所构成的雷达图。

## 参考内容

[] http://blog.csdn.net/qll125596718/article/details/8243404, 2014-11-30