# 谱聚类算法所定义的划分：

聚类问题就是根据相似度，把点分到不同的群组里。如果我们已经知道数据的相似图，那么这个问题事实上是：要找到一个图的划分，使得连接不同群组的边的权重尽可能小（意味着不同类中的点之间是不同的），而在群组内部的边有着很高的权重（意味着相同类中的点是彼此相似的）。这一问题的解决并不容易，而谱聚类算法事实上解决了这一问题的一个放松形式。

已知一个图的邻接矩阵IMG_256以及目标聚类数目kk，要构造一个图的划分，最简单直接的方法就是解决如下的mincutmincut问题：

cut(A1,...,Ak)=1/2⋅∑ki=1W(Ai,Ai¯)

其中IMG_256代表图中除IMG_256以外的点（称为IMG_256的补），IMG_256代表连接IMG_256与IMG_256间点的所有边的权值之和，引入系数 1/21/2则是为了削减重复计算。

在实现中，上述的mincutmincut问题的解决可以是很高效的，但其给出的解作为聚类的结果，常常却非常不理想。一个严重的问题就是，很容易得到单独的点作为一个类。这显然是我们不乐于见到的。

于是我们想办法改善mincutmincut的情况，最常见的两种新的目标函数诞生了：RatioCutRatioCut 和NcutNcut。引入这两个目标函数的动机很简单：控制每个聚类中点的数目都在比较大的水平上。但为了实现这一动机，它们采取的手段还是有着细小的区别。

IMG_256

IMG_256

其中IMG_256代表聚类 AiAi中所有边的权值和。

可以看到，在RatioCutRatioCut中，通过除掉聚类中的顶点数来控制聚类的大小；而在NcutNcut中，则通过除掉聚类中边的权值和来控制聚类的大小。这样的方法确实能够起到作用，为说明这一点，只需注意到式IMG_256和IMG_256的值都是在所有分母取相同值时达到最大即可。

现在的情况看起来很不错，只需要解决RatioCutRatioCut或者NcutNcut就好了，可往下走的话，问题就出现了：原来的mincutmincut是可以高效解的；

我们再也无法像以前解决mincutmincut问题一样简单地解决它了。

这当然抵挡不了学者们执着的心，即使看起来堵死的路，也要把它走出来。

于是一条看起来很自然的路子出现了：把最小化RatioCutRatioCut和NcutNcut的条件进行放松，以达到可以快速求解的目的。

但是，在这之前，为了看清楚问题的本来面目，我们需要转换一下它的表述形式。

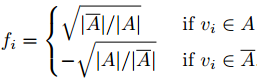
（待会儿可以看到，当我们开始放松条件时，谱聚类算法作为这一问题的一种解决方案开始浮出水面）

现在，以 RatioCutRatioCut 和 聚类数目k=2k=2为例，来解释这个问题。

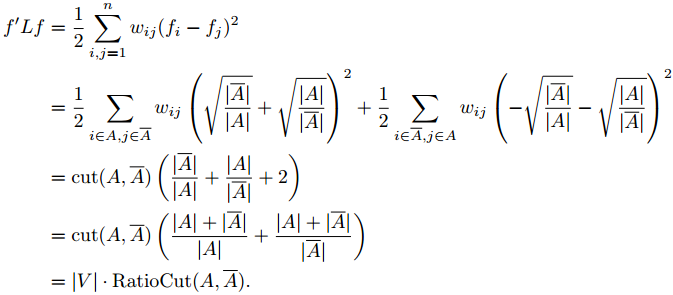
此时，我们的目标是求解最优化问题：

IMG_256

假如我们定义一个向量IMG_256，满足

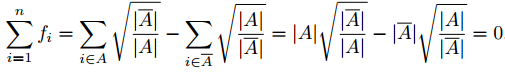
（1）

那么神奇的事情出现了：目标函数RatioCutRatioCut可以通过非标准化LaplacianLaplacian矩阵LL表示出来。具体如下：

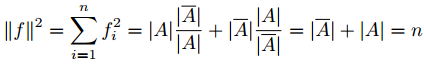


又由于向量 IMG_256满足如下性质：

1. 与常向量1正交，即



1. 范数的平方等于n，即



（3）因此，最小化RatioCutRatioCut的问题被我们转换成了这样的形式：

求 IMG_256 ，满足 ff与常向量11正交、范数等于 n−−√n，且是依照 （1）（1）式定义的向量。

这是一个离散优化问题，因为解向量 ff 的元素只能取两个特定的值。

当然这个问题仍然是NP难的。

关键的时刻到了，我们要对这个问题上面的形式进行放松：把离散的情形转化成连续的情形：

IMG_256

求，满足ff与常向量1正交、范数等于 n−−√n。

现在终于可以快速地解决这一问题了。

通过Rayleigh−RitzRayleigh−Ritz定理，我们知道，非标准化的LaplacianLaplacian阵LL的第二小特征值对应的特征向量正是在上述放松形式下，所要求的向量（注意到L的最小特征值为00，其对应的特征向量为常向量11）。

所以，通过给出矩阵LL的第二个特征向量，我们能给出最小化RatioCutRatioCut的一个近似解。

然而，为了得到图的一个划分，还需要对求得的向量ff进行一定的变换，使得它能指明每个点所属的聚类。

最简单的方法是按照下面的规则来对图进行划分：

IMG_256

不过这不是目标聚类数目k>2k>2时可以沿用的。一般情况下，都把fifi视作RR中的点，用K−meansK−means聚类方法把他们聚成两类，然后按照规则来对图进行划分。

现在已经不难发现，这不就是谱聚类的方法吗？

计算LaplacianLaplacian阵的特征向量，再对它们进行聚类，按照聚类结果进行图的划分。

# 优点：

与传统的聚类算法如k-means算法、层次聚类、DBSCAN算法等相比，谱聚类具有很多优势。谱聚类算法所得到的结果经常优于传统方法，谱聚类实现起来非常简单，可以用标准的线性代数方法高效求解。

# 缺点：

如果降维的幅度不够，效果和效率均不够好；依赖于相似度矩阵。