## 第九章 聚类

9.1 考虑一个由个数据向量组成的数据集，其中每个向量具有32个分量，而每个分量是4字节值。假定向量量化用于压缩，并且使用个原型向量。压缩前后该数据集各需要多少字节的存储空间，压缩率是多少？

参考答案：在压缩之前,数据集需要字节。压缩后的数据集，对于原型向量，需要 字节，对于向量需要 字节，因为识别每个对应数据向量的原型向量需要2个字节。因此，压缩后一共需要10,485,760字节的存储空间，压缩率为12.8。

9.2 影响聚类算法效果的主要原因有哪些：

A、特征选取 B、模式相似性测度 C、分类准则 D、已知类别的样本质量

参考答案：ABC

9.3 如何优化K-means算法？

参考答案： 使用kd树或者ball tree

将所有的观测实例构建成一棵kd树，之前每个聚类中心都是需要和每个观测点做依次距离计算，现在这些聚类中心根据kd树只需要计算附近的一个局部区域即可。

9.4 Kmeans初始类簇中心点的如何选取？

参考答案：

K-means++算法选择初始seeds的基本思想就是：初始的聚类中心之间的相互距离要尽可能的远。

1. 从输入的数据点集合中随机选择一个点作为第一个聚类中心

2. 对于数据集中的每一个点x，计算它与最近聚类中心(指已选择的聚类中心)的距离D(x)

3. 选择一个新的数据点作为新的聚类中心，选择的原则是：D(x)较大的点，被选取作为聚类中心的概率较大

4. 重复2和3直到K个聚类中心被选出来

5. 利用这K个初始的聚类中心来运行标准的K-means算法

9.5 如何评估聚类模型的指标？

参考答案：通过调整兰德系数（adjusted rand score）.

9.6 对于使用K均值对时间序列数据聚类，余弦度量是合适的相似性度量吗？为什么？如果不是，哪种相似性度量更合适？

参考答案： 时间序列数据是密集的高维数据，而余弦度量适合稀疏数据，因此，余弦度量不合适。如果一个时间序列的大小很重要,那么欧几里得距离更合适。如果只有时间序列的形状很重要，那么相关性度量更合适

9.7 常用的聚类划分方式有哪些？列举代表算法。

参考答案：

1. 基于划分的聚类:K-means，k-medoids，CLARANS。

2. 基于层次的聚类：AGNES（自底向上），DIANA（自上向下）。

3. 基于密度的聚类：DBSACN，OPTICS，BIRCH(CF-Tree)，CURE。

4. 基于网格的方法：STING，WaveCluster。

5. 基于模型的聚类：EM,SOM，COBWEB。

9.8 给定具有100个记录的数据集, 要求对数据聚类。使用K均值对数据聚类，但是对于所有的K值（1≤K≤100），K均值算法都只返回一个非空簇。再用K均值的增量版本，但得到的结果完全相同。这怎么回事?用单链或DBSCAN处理该数据，结果如何？

参考答案：

(a)数据完全由一个对象的副本组成。

(b)单链接(和许多其他的凝聚层次聚类方法)会生成一个层次聚类，但哪个点出现在哪个簇取决于点的顺序和具体的算法。如果树图显示每个对象合并的距离，那么数据是由重复的数据组成是显而易见的。DBSCAN会发现所有点都是相互连接的核心点，并生成一个聚类簇。

9.9 传统的凝聚层次聚类过程每步合并两个簇。这样的方法能够正确地捕获数据点集的（嵌套的）簇结构吗？如果不能，解释如何对结果进行后处理，以得到簇结构更正确的视图。

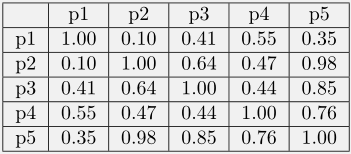
参考答案：

(a)这样的方法不能够准确地捕捉数据点集的簇结构。

例如，考虑一组有三个簇的数据集，每个簇分别有两个、三个和四个子簇。理想的分层聚类的簇结构为：根节点下有三个分支，分别为两个、三个和四个子簇分支，而传统的凝聚层次聚类不能生成这种结构。

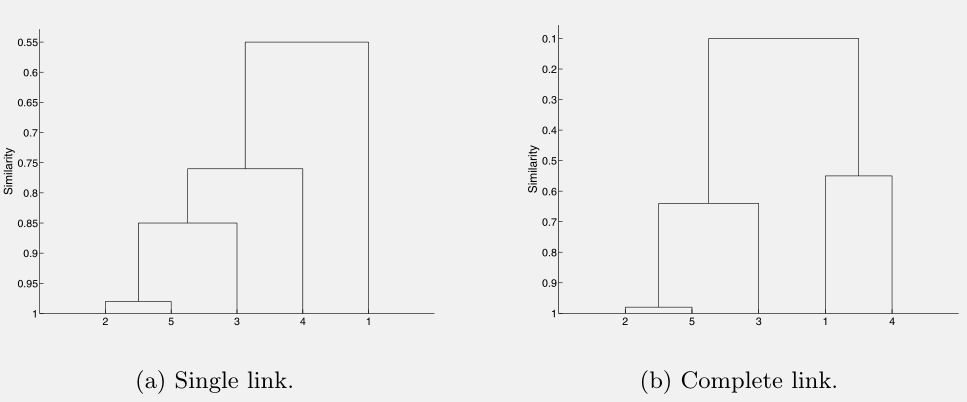
(b)最简单后处理是试图扁平层次聚类.

9.10 使用下表中的相似度矩阵进行单链和全链层次聚类。绘制树状图显示结果。树状图应当清楚地显示合并的次序。



习题10的相似度矩阵

参考答案：解决方案如下图所示



9.11 有时，层次聚类用来产生K个簇，K>1。方法是取树状图的第K层（根在第一层）的簇。通过观察这种方法产生的簇，我们可以评估不同数据和簇类型的层次聚类行为，并且将层次聚类与K均值进行比较。

下面是一维点的集合：{6，12，18，24，30，42，48}。

（a）对于下列每组初始质心，将每个点指派到最近的质心，创建两个簇，然后对两个簇的每组质心分别计算总平方误差，对每组质心，给出这两个簇和总平方误差。

（1）{18,45}

（2）{15,40}

（b）两组质心代表稳定解吗，即如果在该数据集上，使用给定的质心作为初始质心运动K均值，所产生的簇会有所改变吗？

（c）单链产生的簇是什么？

（d）在此情况下，哪种技术（K均值或单链）能够产生“最自然的”簇？（对于K均值，用最小平方误差产生聚类。）

（e）这个自然聚类对应于哪种簇定义？（明显分离的、基于中心的、基于邻近的或基于密度的。）

（f）K均值算法的哪个著名特征解释了前面的行为？

参考答案：

(a)i. {18, 45}

第一个簇为{6, 12, 18, 24, 30}

平方误差 = 360.

第二个簇为{42,48}

平方误差 = 18

总平方误差 = 378

ii. {15, 40}

第一个簇为{6, 12, 18, 24}

平方误差 = 180

第二个簇为{30, 42, 48}

平方误差 = 168

总平方误差 = 348.

(b)是

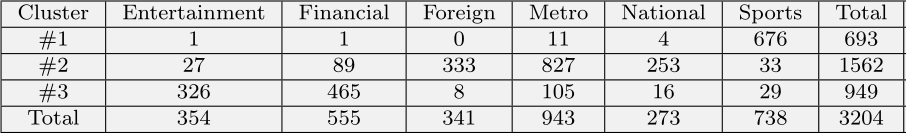
(c)两个簇分别为 { 6、12、18、24、30 }和{ 42,48 }

(d) MIN 技术生成”最自然的”簇

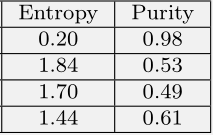
(e) MIN 生成邻近的聚类簇。基于密度也是可选的答案，甚至基于中心也是可选的，因为一组中心点提供所需的簇。

(f) K均值不适用于寻找大小不同的簇，特别是当它们没有很好的分离时。原因是优化目标为最小化平方误差时会“破坏”更大的簇。因此，在这个问题中，低误差聚类方法是“非自然”的。

9.12 计算下表的混淆矩阵的熵和纯度



参考答案：



9.13 使用下一题的数据计算每个点、每个簇和整个聚类的轮廓系数。

参考答案：

簇1包含{ P1,P2 },簇2包含{ P3,P4 }，我们从相似矩阵获得的相异度矩阵如下：

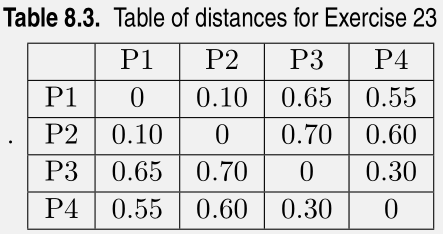


Table of distances

令a表示一个点到它所在簇的其它点的平均距离

令b表示一个点到其它簇中点的平均距离的最小值

Point P1: SC = 1- a/b = 1 - 0.1/((0.65+0.55)/2)= 5/6 = 0.833

Point P2: SC = 1- a/b = 1 - 0.1/((0.7+0.6)/2) = 0.846

Point P2: SC = 1- a/b = 1 - 0.3/((0.65+0.7)/2) = 0.556

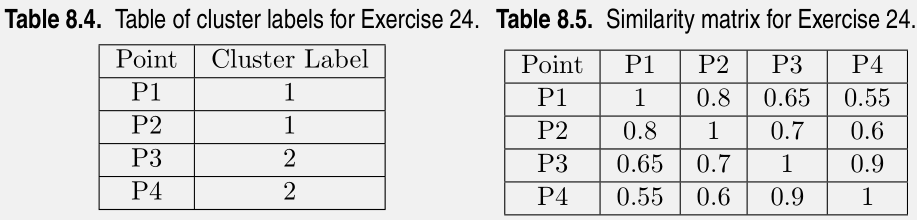
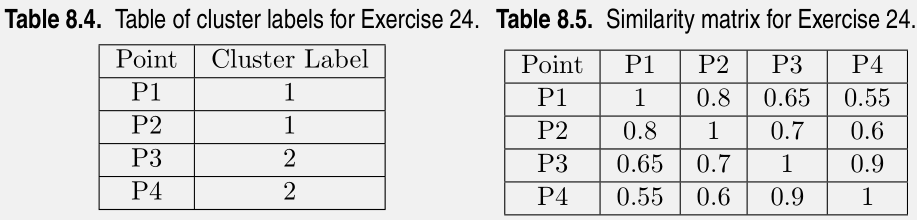
Point P2: SC = 1- a/b = 1 - 0.3/((0.55+0.6)/2) = 0.478

Cluster 1 Average SC = (0.833+0.846)/2 = 0.84

Cluster 2 Average SC = (0.556+0.478)/2 = 0.52

Overall Average SC = (0.840+0.517)/2 = 0.68

9.14 给定分别由下表给出的簇标号集和相似度矩阵，计算该相似度矩阵与理想的相似度矩阵之间的相关度。如果两个对象属于同一个簇，理想的相似度矩阵的第ij项为1，否则为0.

簇标号表 相似度矩阵

参考答案：

我们需要计算向量x和y之间的相关性，x =< 1,0,0,0,0,1 >，y =< 0.8,0.65,0.55,0.7,0.6,0.6>

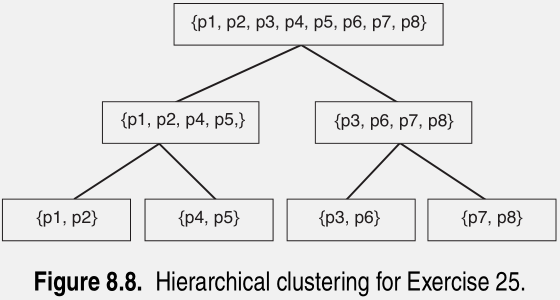
向量x的标准差为 : σx = 0.5164

向量y的标准差为 : σy = 0.1703

协方差为 : cov(x,y) = −0.200

相关系数 = cov(x,y)/σxσ y = −0.227

9.15 对于8个对象{p1,p2,p3,p4,p5,p6,p7,p8}和下图显示的层次聚类，计算层次F度量。类A包含点p1、p2和p3，而p4、p5、p6、p7和p8属于类B。



习题15的层次聚类

参考答案：

令表示类别i与簇j之间的召回率

令表示类别i与簇j之间的准确率

表示类别i与簇j的F度量值

对于簇 #1 = {p1,p2,p3,p4,p5,p6,p7,p8}:

Class = A:

R(A,1) = 3/3 = 1， P(A,1) = 3/8 = 0.375

F(A,1) = 2 × 1 × 0.375/(1 + 0.375) = 0.55

Class = B:

R(B,1) = 5/5 = 1， P(A,1) = 5/8 = 0.625， F(A,1) = 0.77

对于簇 #2= {p1,p2,p4,p5}

Class = A:

R(A,2) = 2/3，P(A,2) = 2/4,F(A,2) = 0.57

Class = B:

R(B,2) = 2/5，P(B,2) = 2/4，F(B,2) = 0.44

对于簇 #3= {p3, p6, p7, p8}

Class = A:

R(A,3) = 1/3，P(A,3) = 1/4，F(A,3) = 0.29

Class =B:

R(B,3) = 3/5，P(B,3) = 3/4，F(B,3) = 0.67

对于簇 #4={p1, p2}

Class = A:

R(A,4) = 2/3，P(A,4) = 2/2， F(A,4) = 0.8

Class =B:

R(B,4) = 0/5，P(B,4) = 0/2， F(B,4) = 0

对于簇 #5 = {p4, p5}

Class = A:

R(A,5) = 0，P(A,5) = 0， F(A,5) = 0

Class =B:

R(B,5) = 2/5， P(B,5) = 2/2， F(B,5) = 0.57

对于簇 #6 = {p3, p6}

Class = A:

R(A,6) = 1/3， P(A,6) = 1/2， F(A,6) = 0.4

Class =B:

R(B,6) = 1/5， P(B,6) = 1/2，F(B,6) = 0.29

对于簇 #7 = {p7, p8}

Class = A:

R(A,7) = 0， P(A,7) = 1， F(A,7) = 0

Class = B:

R(B,7) = 2/5， P(B,7) = 2/2， F(B,7) = 0.57

Class A: F(A) = max{F(A,j)} = max{0.55,0.57,0.29,0.8,0,0.4,0} = 0.8

Class B: F(B) = max{F(B,j)} = max{0.77,0.44,0.67,0,0.57,0.29,0.57} =0.77

全部簇: