

K-means 聚类

一、摘要

分类作为一种监督学习方法，要求必须事先明确知道各个类别的信息，并且断言所有待分类项都有一个类别与之对应。但是很多时候上述条件得不到满足，尤其是在处理海量数据的时候，如果通过预处理使得数据满足分类算法的要求，则代价非常大，这时候可以考虑使用聚类算法。聚类属于无监督学习，相比于分类，聚类不依赖预定义的类和类标号的训练实例。本文首先介绍聚类的基础——距离与相异度，然后介绍一种常见的聚类算法——k 均值和 k 中心点聚类，最后会举一个实例：应用聚类方法试图解决一个在体育界大家颇具争议的问题——中国男足近几年在亚洲到底处于几流水平。

二、相异度计算

在正式讨论聚类前，我们要先弄清楚一个问题：如何定量计算两个可比较元素间的相异度。用通俗的话说，相异度就是两个东西差别有多大，例如人类与章鱼的相异度明显大于人类与黑猩猩的相异度，这是能我们直观感受到的。但是，计算机没有这种直观感受能力，我们必须对相异度在数学上进行定量定义。

设 $X = \{x_1, x_2, \dots, x_n\}, Y = \{y_1, y_2, \dots, y_n\}$ ，其中 X, Y 是两个元素项，各自具有 n 个可度量特征属性，那么 X 和 Y 的相异度定义为： $d(X, Y) = f(X, Y) \rightarrow R$ ，其中 R 为实数域。也就是说相异度是两个元素对实数域的一个映射，所映射的实数定量表示两个元素的相异度。

下面介绍不同类型变量相异度计算方法。

1、标量

标量也就是无方向意义的数字，也叫标度变量。现在先考虑元素的所有特征属性都是标量的情况。例如，计算 $X=\{2,1,102\}$ 和 $Y=\{1,3,2\}$ 的相异度。一种很自然的想法是用两者的欧几里得距离来作为相异度，欧几里得距离的定义如下：

$$d(X, Y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \dots + (x_n - y_n)^2}$$

其意义就是两个元素在欧氏空间中的集合距离，因为其直观易懂且可解释性强，被广泛用于标识两个标量元素的相异度。将上面两个示例数据代入公式，可得两者的欧氏距离为：

$$d(X, Y) = \sqrt{(2 - 1)^2 + (1 - 3)^2 + (102 - 2)^2} = 100.025$$

除欧氏距离外，常用作度量标量相异度的还有曼哈顿距离和闵可夫斯基距离，两者定义如下：

曼哈顿距离：
$$d(X, Y) = |x_1 - y_1| + |x_2 - y_2| + \dots + |x_n - y_n|$$

闵可夫斯基距离：
$$d(X, Y) = \sqrt[p]{|x_1 - y_1|^p + |x_2 - y_2|^p + \dots + |x_n - y_n|^p}$$

欧氏距离和曼哈顿距离可以看做是闵可夫斯基距离在 $p=2$ 和 $p=1$ 下的特例。另外这三种距离都可以加权，这个很容易理解，不再赘述。

下面要说一下标量的规格化问题。上面这样计算相异度的方式有一点问题，就是取值范围大的属性对距离的影响高于取值范围小的属性。例如上述例子中第三个属性的取值跨度远

大于前两个，这样不利于真实反映真实的相异度，为了解决这个问题，一般要对属性值进行规格化。所谓规格化就是将各个属性值按比例映射到相同的取值区间，这样是为了平衡各个属性对距离的影响。通常将各个属性均映射到[0,1]区间，映射公式为：

$$a'_i = \frac{a_i - \min(a_i)}{\max(a_i) - \min(a_i)}$$

其中 $\max(a_i)$ 和 $\min(a_i)$ 表示所有元素项中第 i 个属性的最大值和最小值。例如，将示例中的元素规格化到[0,1]区间后，就变成了 $X'=\{1,0,1\}$ ， $Y'=\{0,1,0\}$ ，重新计算欧氏距离约为 1.732。

2、二元变量

所谓二元变量是只能取 0 和 1 两种值变量，有点类似布尔值，通常用来标识是或不是这种二值属性。对于二元变量，上一节提到的距离不能很好标识其相异度，我们需要一种更适合的标识。一种常用的方法是用元素相同序位同值属性的比例来标识其相异度。

设有 $X=\{1,0,0,1,0,1,1\}$ ， $Y=\{0,0,0,1,1,1,1\}$ ，可以看到，两个元素第 2、3、5、7 和 8 个属性取值相同，而第 1、4 和 6 个取值不同，那么相异度可以标识为 $3/8=0.375$ 。一般的，对于二元变量，相异度可用“取值不同的同位属性数/单个元素的属性位数”标识。

上面所说的相异度应该叫做对称二元相异度。现实中还有一种情况，就是我们只关心两者都取 1 的情况，而认为两者都取 0 的属性并不意味着两者更相似。例如在根据病情对病人聚类时，如果两个人都患有肺癌，我们认为两个人增强了相似度，但如果两个人都没患肺癌，并不觉得这加强了两人的相似性，在这种情况下，改用“取值不同的同位属性数/(单个元素的属性位数-同取 0 的位数)”来标识相异度，这叫做非对称二元相异度。如果用 1 减去非对称二元相异度，则得到非对称二元相似度，也叫 Jaccard 系数，是一个非常重要的概念。

3、分类变量

分类变量是二元变量的推广，类似于程序中的枚举变量，但各个值没有数字或序数意义，如颜色、民族等等，对于分类变量，用“取值不同的同位属性数/单个元素的全部属性数”来标识其相异度。

4、序数变量

序数变量是具有序数意义的分类变量，通常可以按照一定顺序意义排列，如冠军、亚军和季军。对于序数变量，一般为每个值分配一个数，叫做这个值的秩，然后以秩代替原值当做标量属性计算相异度。

5、向量

对于向量，由于它不仅有大而且方向，所以闵可夫斯基距离不是度量其相异度的好办法，一种流行的做法是用两个向量的余弦度量，其度量公式为：

$$s(X, Y) = \frac{X'Y}{\|X\| \|Y\|}$$

其中 $\|X\|$ 表示 X 的欧几里得范数。要注意，余弦度量度量的不是两者的相异度，而是相似度！

三、聚类问题

在讨论完了相异度计算的问题，就可以正式定义聚类问题了。

所谓聚类问题，就是给定一个元素集合 **D**，其中每个元素具有 **n** 个可观察属性，使用某种算法将 **D** 划分成 **k** 个子集，要求每个子集内部的元素之间相异度尽可能低，而不同子集的元素相异度尽可能高。其中每个子集叫做一个簇。

与分类不同，分类是示例式学习，要求分类前明确各个类别，并断言每个元素映射到一个类别，而聚类是观察式学习，在聚类前可以不知道类别甚至不给定类别数量，是无监督学习的一种。目前聚类广泛应用于统计学、生物学、数据库技术和市场营销等领域，相应的算法也非常的多。本文仅介绍一种最简单的聚类算法——**k** 均值 (**k-means**) 算法。

四、K-means 算法及其示例

k 均值算法的计算过程非常直观：

1、从 **D** 中随机取 **k** 个元素，作为 **k** 个簇的各自的中心。

2、分别计算剩下的元素到 **k** 个簇中心的相异度，将这些元素分别划归到相异度最低的簇。

3、根据聚类结果，重新计算 **k** 个簇各自的中心，计算方法是取簇中所有元素各自维度的算术平均数。

4、将 **D** 中全部元素按照新的中心重新聚类。

5、重复第 4 步，直到聚类结果不再变化。

6、将结果输出。

由于算法比较直观，没有什么可以过多讲解的。下面，我们来看看 **k-means** 算法一个有趣的应用示例：中国男足近几年到底在亚洲处于几流水平？

今年中国男足可算是杯具到家了，几乎到了过街老鼠人人喊打的地步。对于目前中国男足在亚洲的地位，各方也是各执一词，有人说中国男足亚洲二流，有人说三流，还有人说根本不入流，更有人说其实不比日韩差多少，是亚洲一流。既然争论不能解决问题，我们就让数据告诉我们结果吧。

下图是我采集的亚洲 15 只球队在 2005 年-2010 年间大型杯赛的战绩（由于澳大利亚是后来加入亚足联的，所以这里没有收录）。

	A	B	C	D
1		2006年世界杯	2010年世界杯	2007年亚洲杯
2	中国	50	50	9
3	日本	28	9	4
4	韩国	17	15	3
5	伊朗	25	40	5
6	沙特	28	40	2
7	伊拉克	50	50	1
8	卡塔尔	50	40	9
9	阿联酋	50	40	9
10	乌兹别克斯坦	40	40	5
11	泰国	50	50	9
12	越南	50	50	5
13	阿曼	50	50	9
14	巴林	40	40	9
15	朝鲜	40	32	17
16	印尼	50	50	9

其中包括两次世界杯和一次亚洲杯。我提前对数据做了如下预处理：对于世界杯，进入决赛圈则取其最终排名，没有进入决赛圈的，打入预选赛十强赛赋予 40，预选赛小组未出线的赋予 50。对于亚洲杯，前四名取其排名，八强赋予 5，十六强赋予 9，预选赛没出现的赋予 17。这样做是为了使得所有数据变为标量，便于后续聚类。

下面先对数据进行[0,1]规格化，下面是规格化后的数据：

	A	B	C	D
1		2006年世界杯	2010年世界杯	2007年亚洲杯
2	中国	1	1	0.5
3	日本	0.3	0	0.19
4	韩国	0	0.15	0.13
5	伊朗	0.24	0.76	0.25
6	沙特	0.3	0.76	0.06
7	伊拉克	1	1	0
8	卡塔尔	1	0.76	0.5
9	阿联酋	1	0.76	0.5
10	乌兹别克斯坦	0.7	0.76	0.25
11	泰国	1	1	0.5
12	越南	1	1	0.25
13	阿曼	1	1	0.5
14	巴林	0.7	0.76	0.5
15	朝鲜	0.7	0.68	1
16	印尼	0.7	0.68	0.5

接着用 k-means 算法进行聚类。设 k=3，即将这 15 支球队分成三个集团。

现抽取日本、巴林和泰国的值作为三个簇的种子，即初始化三个簇的中心为 A: {0.3, 0, 0.19}, B: {0.7, 0.76, 0.5} 和 C: {1, 1, 0.5}。下面，计算所有球队分别对三个中心点的相异度，这里以欧氏距离度量。下面是我用程序求取的结果：

file:///E:/个人文档/kmeans/kmeans/bin/Debug/kmeans.EXE			
1.212436	0.519615	0	
0	0.69282	1.212436	
0.519615	1.212436	1.732051	
0.103923	0.796743	1.316359	
0	0.69282	1.212436	
1.212436	0.519615	0	
1.212436	0.519615	0	
1.212436	0.519615	0	
0.69282	0	0.519615	
1.212436	0.519615	0	
1.212436	0.519615	0	
1.212436	0.519615	0	
0.69282	0	0.519615	
0.69282	0	0.519615	
1.212436	0.519615	0	

从做到右依次表示各支球队到当前中心点的欧氏距离，将每支球队分到最近的簇，可对各支球队做如下聚类：

中国 C，日本 A，韩国 A，伊朗 A，沙特 A，伊拉克 C，卡塔尔 C，阿联酋 C，乌兹别

克斯坦 B, 泰国 C, 越南 C, 阿曼 C, 巴林 B, 朝鲜 B, 印尼 C。

第一次聚类结果:

A: 日本, 韩国, 伊朗, 沙特;

B: 乌兹别克斯坦, 巴林, 朝鲜;

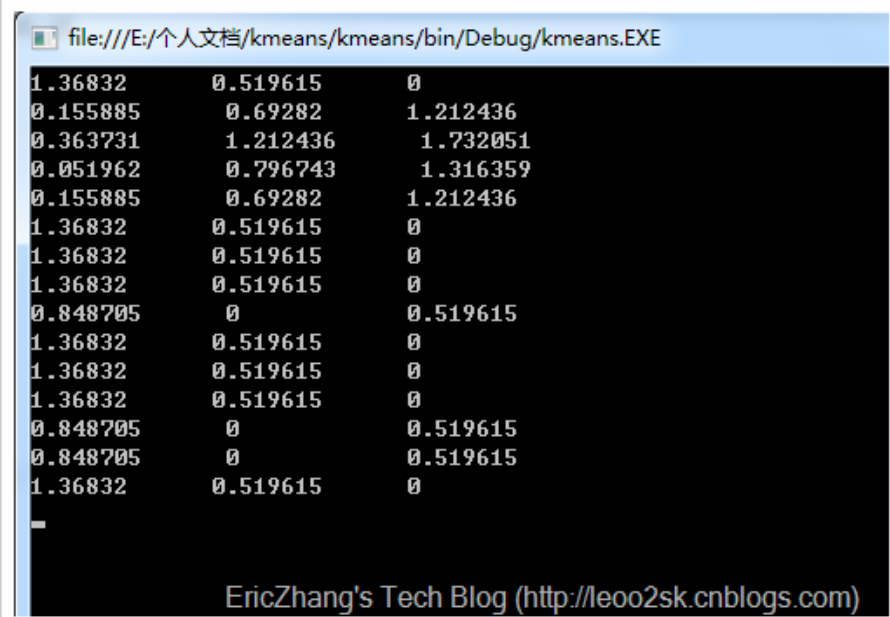
C: 中国, 伊拉克, 卡塔尔, 阿联酋, 泰国, 越南, 阿曼, 印尼。

下面根据第一次聚类结果, 调整各个簇的中心点。

A 簇的新中心点为: $\{(0.3+0+0.24+0.3)/4=0.21, (0+0.15+0.76+0.76)/4=0.4175, (0.19+0.13+0.25+0.06)/4=0.1575\} = \{0.21, 0.4175, 0.1575\}$

用同样的方法计算得到 B 和 C 簇的新中心点分别为 $\{0.7, 0.7333, 0.4167\}$, $\{1, 0.94, 0.40625\}$ 。

用调整后的中心点再次进行聚类, 得到:



```
file:///E:/个人文档/kmeans/kmeans/bin/Debug/kmeans.EXE
1.36832      0.519615      0
0.155885     0.69282       1.212436
0.363731     1.212436      1.732051
0.051962     0.796743      1.316359
0.155885     0.69282       1.212436
1.36832      0.519615      0
1.36832      0.519615      0
1.36832      0.519615      0
0.848705     0             0.519615
1.36832      0.519615      0
1.36832      0.519615      0
1.36832      0.519615      0
0.848705     0             0.519615
0.848705     0             0.519615
1.36832      0.519615      0

EricZhang's Tech Blog (http://leoo2sk.cnblogs.com)
```

第二次迭代后的结果为:

中国 C, 日本 A, 韩国 A, 伊朗 A, 沙特 A, 伊拉克 C, 卡塔尔 C, 阿联酋 C, 乌兹别克斯坦 B, 泰国 C, 越南 C, 阿曼 C, 巴林 B, 朝鲜 B, 印尼 C。

结果无变化, 说明结果已收敛, 于是给出最终聚类结果:

亚洲一流: 日本, 韩国, 伊朗, 沙特

亚洲二流: 乌兹别克斯坦, 巴林, 朝鲜

亚洲三流: 中国, 伊拉克, 卡塔尔, 阿联酋, 泰国, 越南, 阿曼, 印尼

看来数据告诉我们, 说国足近几年处在亚洲三流水平真的是没有冤枉他们, 至少从国际杯赛战绩是这样的。

其实上面的分析数据不仅告诉了我们聚类信息, 还提供了一些其它有趣的信息, 例如从中可以定量分析出各个球队之间的差距, 例如, 在亚洲一流队伍中, 日本与沙特水平最接近, 而伊朗则相距他们较远, 这也和近几年伊朗没落的实际相符。另外, 乌兹别克斯坦和巴林虽然没有打进近两届世界杯, 不过凭借预选赛和亚洲杯上的出色表现占据 B 组一席之地, 而朝鲜由于打入了 2010 世界杯决赛圈而有幸进入 B 组, 可是同样奇迹般夺得 2007 年亚洲杯的伊拉克却被分在三流, 看来亚洲杯冠军的分量还不如打进世界杯决赛圈重啊。其它有趣的信息, 有兴趣的朋友可以进一步挖掘。

出处: <http://www.cnblogs.com/leoo2sk/archive/2010/09/20/k-means.html>

其他阐述：

1、聚类算法之 K-means

区分两个概念：

hard clustering：一个文档要么属于类 w ，要么不属于类 w ，即文档对确定的类 w 是二值的 1 或 0。

soft clustering：一个文档可以属于类 w_1 ，同时也可以属于 w_2 ，而且文档属于一个类的值不是 0 或 1，可以是 0.3 这样的小数。

K-Means 就是一种 **hard clustering**，所谓 **K-means** 里的 **K** 就是我们要事先指定分类的个数，即 **K** 个。

k-means 算法的流程如下：

- 1) 从 N 个文档随机选取 K 个文档作为初始质心
- 2) 对剩余的每个文档测量其到每个质心的距离，并把它归到最近的质心的类
- 3) 重新计算已经得到的各个类的质心
- 4) 迭代 2~3 步直至满足既定的条件，算法结束

在 **K-means** 算法里所有的文档都必须向量化， n 个文档的质心可以认为是这 n 个向量的中心，计算方法如下：

$$\bar{\mu}(\omega) = \frac{1}{|\omega|} \sum_{\vec{x} \in \omega} \vec{x}$$

这里加入一个方差 **RSS** 的概念：

$$RSS_k = \sum_{\vec{x} \in \omega_k} |\vec{x} - \bar{\mu}(\omega_k)|^2$$

$$RSS = \sum_{k=1}^K RSS_k$$

RSS_k 的值是类 k 中每个文档到质心的距离，**RSS** 是所有 k 个类的 **RSS** 值的和。

算法结束条件：

- 1) 给定一个迭代次数，达到这个次数就停止，这好像不是一个好建议。
- 2) k 个质心应该达到收敛，即第 n 次计算出的 n 个质心在第 $n+1$ 次迭代时候位置不变。
- 3) n 个文档达到收敛，即第 n 次计算出的 n 个文档分类和在第 $n+1$ 次迭代时候文档分类结果相同。

- 4) **RSS** 值小于一个阈值，实际中往往把这个条件结合条件 1 使用

回过头用 **RSS** 讨论质心的计算方法是否合理

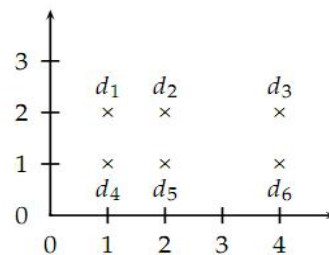
$$RSS_k(\vec{v}) = \sum_{\vec{x} \in \omega_k} |\vec{v} - \vec{x}|^2 = \sum_{\vec{x} \in \omega_k} \sum_{m=1}^M (v_m - x_m)^2$$
$$\frac{\partial RSS_k(\vec{v})}{\partial v_m} = \sum_{\vec{x} \in \omega_k} 2(v_m - x_m)$$

为了取得 **RSS** 的极小值，**RSS** 对质心求偏导数应该为 0，所以得到质心

$$\bar{\mu}(\omega) = \frac{1}{|\omega|} \sum_{\vec{x} \in \omega} \vec{x}$$

可见，这个质心的选择是合乎数学原理的。

K-means 方法的缺点是聚类结果依赖于初始选择的几个质点位置，看下面这个例子：



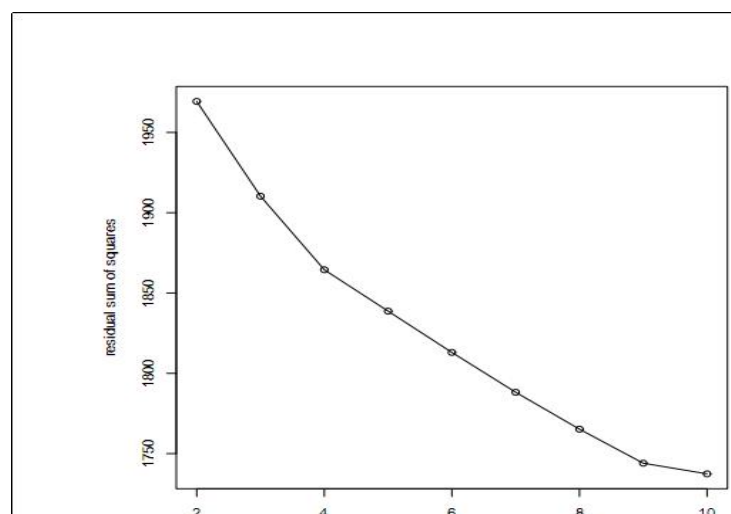
如果使用 2-means 方法，初始选择 d2 和 d5 那么得到的聚类结果就是 {d1, d2, d3} {d4, d5, d6}，这不是一个合理的聚类结果

解决这种初始种子问题的方案：

- 1) 去处一些游离在外层的文档后再选择
- 2) 多选一些种子，取结果好的 (RSS 小) 的 K 个类继续算法
- 3) 用层次聚类的方法选择种子。我认为这不是一个合适的方法，因为对初始 N 个文档进行层次聚类代价非常高。

以上的讨论都是基于 K 是已知的，但是我们怎么能从随机的文档集合中选择这个 k 值呢？

我们可以对 k 去 1~N 分别执行 k-means，得到 RSS 关于 K 的函数下图：



当 RSS 由显著下降到不是那么显著下降的 K 值就可以作为最终的 K，如图可以选择 4 或 9。

作者 Email: [luoleicn\(at\)gmail.com](mailto:luoleicn(at)gmail.com)

<http://blog.csdn.net/luoleicn/archive/2010/03/05/5350625.aspx>

2、k-均值聚类算法 (k-means)

K-均值聚类(K-means clustering)是 Mac Queen 提出的一种非监督实时聚类算法，在最小化误差函数的基础上将数据划分为预定的类数 K。该算法原理简单并便于处理大量数据，在基因表达数据分析中得到广泛应用，如 Tavazoie 等应用 K-means 聚类酵母细胞周期表达数据。在 K-means 算法运行前必须先指定聚类数目 K 和迭代次数或收敛条件，并指定 K 个初始聚类中心，根据一定的相似性度量准则，将每一条基因分配到最近或“相似”的聚类中心，形成类，然后以每一类的平均矢量作为这一类的聚类中心，重新分配，反复迭代直到类收敛或达到最大的迭代次数。

K-means 聚类算法对初始聚类中心依赖性比较大, 随机选取初始聚类中心的缺点是如果使得初始聚类中心得到的分类严重偏离全局最优分类, 这样算法可能会陷入局部最优值。而且当聚类数比较大的时候, 这种缺点更为明显, 往往要经过多次聚类才有可能达到较满意的结果。**Yeung** 等提出了采用均连接层次聚类结果初始化 **K-means** 聚类中心。此方法有效地排除了随机初始化过程中引入的随机性因素, 使得算法成为确定性的, 可以得到稳定的聚类结果; 而且, 这种初始化方式也能够利用数据中的类结构信息, 使得聚类质量相对于随机初始化时的平均质量有显著的提高。

K-means 聚类算法的一般步骤:

- 1) 初始化。输入基因表达矩阵作为对象集 **X**, 输入指定聚类类数 **N**, 并在 **X** 中随机选取 **N** 个对象作为初始聚类中心。设定迭代中止条件, 比如最大循环次数或者聚类中心收敛误差容限。
- 2) 进行迭代。根据相似度准则将数据对象分配到最接近的聚类中心, 从而形成一类。初始化隶属度矩阵。
- 3) 更新聚类中心。然后以每一类的平均向量作为新的聚类中心, 重新分配数据对象。
- 4) 反复执行第二步和第三步直至满足中止条件。

该算法理论严密, 实现简单, 已成为很多其它改进算法的基础, 但它对初始码书的选择非常敏感。

出处: <http://apps.hi.baidu.com/share/detail/12920070>