链滴

数据挖掘算法初窥门庭--聚类

作者：Zing

原文链接：https://ld246.com/article/1457699700193

来源网站：链滴

#聚类（Cluster）
##概念
什么是聚类：
按照个体或样品的特征将它们分类，使同一类别的个体具有尽可能高的同质性，而类别之间则应该具尽可能高的异质性。
聚类的特点：
不是一种统计方法，而是数据处理技术；需要自定聚类变量以及类别个数，属于非监督的分析方法；般不涉及有关统计量的分布；不需要进行显著性检验；聚类算法比距离算法对结果影响更大；样本的序会影响聚类的结果。
一些重要概念：

聚类变量：一组表示个体特征的变量。完全由研究者规定，会对结果产生较大的影响。需要需要变量进行标准化处理。
类别个数：聚类结果类别的个数。完全由研究者规定，不管实际数据中是否存在不同的类别吗，能得到若干类别的解。
个体同质程度：有两种方式进行测量

采用描述个体之间的接近程度指标（数量），如“距离”：欧式距离、 曼哈顿距离等
采用描述铬铁之间的相似程度指标（模式），如“相关系数”：皮尔逊相关系数

<hr>
##快速聚类和两阶段聚类
根据聚类算法的处理过程可以分为：快速聚类和两阶段聚类。

快速聚类：

思想：
选取 k 个观测量作为初始聚类中心，以距离最小原则将样本分配到 k 个类中，在每个类中以一定的方重新选举聚类中心。不断迭代，直到收敛或满足要求。
特点：
一般只能处理数值型的变量；噪声对结果影响比较大；强线性关系变量会导致重复贡献影响结果。</l
>

两阶段聚类：

思想：
将聚类分为预聚类（类数目增加）和聚类（类数目减少）两个阶段，类似于，构造一棵树，从根向上断生长出更多的分支，然后对树进行修剪，把小的分支处理掉，合并留下的大分支。
特点：
可以处理数值型和分类型的变量；自动确定最优聚类数目；诊断离群点和噪声数据；可缩放性强。</l
>

<hr>
##常见算法
根据算法思想可以把聚类算法分为下列类型。下面我们简单学习各种算法的思想，特点，优缺点等。体算法在实践中再进行具体的学习。
###划分算法
思想：划分算法属于快速聚类的方法。
1.选取 k 个观测量作为初始聚类中心

2.以距离最小原则将每个实例分配到 k 个类中
3.在每个类中以一定的方法重新选举聚类中心
4.不断迭代，直到收敛或满足要求

k-means 算法

特点：初始类中心选取是任意的，类中心的再选举采用类中所有对象的均值。
优点：算法简单，也是最常用的聚类算法；对大数据是可伸缩的且效率高，时间复杂度接近于线。
缺点：初始值的选取会对结果产生较大的影响；对脏数据很敏感；只能处理数值类型数据

k-medoids 算法

特点：是对 K-MEANS 算法的改进，类中心（medoids）的再选举采用的是选取到类中其他点离之和最小的点。
优点：对脏数据不敏感
缺点：选取类中心计算量大，一般只能用于小数据集

clara 算法

特点：是 k-medoids 效率不好的解决方案，在选举类中心时，使用抽样数据代替整个数据集

优点：提高选举类中心的效率
缺点：效率取决于采样的大小，采样大小决定了聚类的结果，一般不太可能得到最佳结果

clarans

特点：是对 clara 的改进：clara 在选举类中心时是用的采用是不变的，clarans 算法在没一次迭使用的采样都是不一样的。
优点：解决 clara 算法无法得到最佳结果的问题
缺点：必须人为限定迭代次数。

<table>
<tbody><tr>
<td>k-means</td>
<td>是一种典型的划分聚类算法，它用一个聚类的中心来代表一个簇，即在迭代过程中选择的聚点一定是聚类中的一个点，该算法只能处理数值型数据</td>
</tr>
<tr>
<td>k-modes</td>
<td>K-Means算法的扩展，能够处理分类数据，采用简单匹配方法来度量分类型数据的相似度</td>
</tr>
<tr>
<td>k-prototypes</td>
<td>结合了K-Means和K-Modes两种算法，能够处理混合型数据</td>

</tr>
<tr>
<td>k-medoids</td>
<td>在迭代过程中选择簇中到其他点距离之和最小的点为中心，PAM是典型的k-medoids算法</td>
</tr>
<tr>
<td>CLARA</td>
<td>CLARA算法在PAM的基础上采用了抽样技术，能够处理大规模数据</td>
</tr>
<tr>
<td>CLARANS</td>
<td>CLARANS算法融合了PAM和CLARA两者的优点，是第一个用于空间数据库的聚类算法, 该算法用于处理数值型数据</td>
</tr>
</tbody></table>
<hr>
###层次算法
层次聚类方法是对给定数据集进行层次分解明知道某种条件满足为止。具体可以分为凝聚和分裂两种案。

凝聚：自底向上，首先每个对象作为一簇，然后合并这些原子簇，知道某个条件被满足。
分裂：自顶向下，首先将所有的对象置于同一个簇，然后逐渐分裂为更小的簇，直到某个条件被足。

<table>
<tbody><tr>
<td>CURE</td>
<td>采用抽样技术先对数据集随机抽取样本，再采用分区技术对样本进行分区，然后对每个分区局聚类，最后对局部聚类进行全局聚类。适合处理数值型数据类型</td>
</tr>
<tr>
<td>ROCK</td>
<td>也采用了随机抽样技术，该算法在计算两个对象的相似度时，同时考虑了周围对象的影响，适处理混合型数据类型</td>
</tr>
<tr>
<td>CHEMALOEN（变色龙算法）</td>
<td>首先由数据集构造成一个K-最近邻图Gk ,再通过一个图的划分算法将图Gk 划分成大量的子图,每子图代表一个初始子簇,最后用一个凝聚的层次聚类算法反复合并子簇，找到真正的结果簇</td>
</tr>
<tr>
<td>SBAC</td>
<td>SBAC算法则在计算对象间相似度时，考虑了属性特征对于体现对象本质的重要程度，对于更能现对象本质的属性赋予较高的权值</td>
</tr>
<tr>
<td>BIRCH</td>
<td>BIRCH算法利用树结构对数据集进行处理，叶结点存储一个聚类，用中心和半径表示，顺序处每一个对象，并把它划分到距离最近的结点，该算法也可以作为其他聚类算法的预处理过程。适合处数值型数据类型</td>
</tr>
<tr>
<td>BUBBLE</td>
<td>BUBBLE算法则把BIRCH算法的中心和半径概念推广到普通的距离空间</td>

</tr>
<tr>
<td>BUBBLE-FM</td>
<td>BUBBLE-FM算法通过减少距离的计算次数，提高了BUBBLE算法的效率</td>
</tr>
</tbody></table>
<hr>
###密度算法
基于距离的算法只能发现“类圆形”的聚类，基于密度的算法克服了这个缺点。
密度算法的知道思想是，当一个区域中的点的密度大于某个阈值，就把它加入到与之相近的聚类中去

<table>
<tbody><tr>
<td>DBSCAN</td>
<td>采用空间索引技术来搜索对象的邻域，引入了“核心对象”和“密度可达”等概念，从核心对出发，把所有密度可达的对象组成一个簇,适合处理数值型数据类型</td>
</tr>
<tr>
<td>GDBSCAN</td>
<td>算法通过泛化DBSCAN算法中邻域的概念，以适应空间对象的特点</td>
</tr>
<tr>
<td>OPTICS</td>
<td>OPTICS算法结合了聚类的自动性和交互性，先生成聚类的次序，可以对不同的聚类设置不同的数，来得到用户满意的结果</td>
</tr>
<tr>
<td>FDC</td>
<td>FDC算法通过构造k-d tree把整个数据空间划分成若干个矩形空间，当空间维数较少时可以大大高DBSCAN的效率</td>
</tr>
</tbody></table>
<hr>
###网格算法
基于网格的算法先将数据空间划分为有限个单元的网格结构，所有的处理都是以单个的单元为对象。格算法的特点是处理速度很快，通常于单元个数有关而与记录个数无关。
<table>
<tbody><tr>
<td>STING</td>
<td>利用网格单元保存数据统计信息，从而实现多分辨率的聚类</td>
</tr>
<tr>
<td>WaveCluster</td>
<td>在聚类分析中引入了小波变换的原理，主要应用于信号处理领域。只能处理数值型数据类型</t
>
</tr>
<tr>
<td>CLIQUE</td>
<td>是一种结合了网格和密度的聚类算法,适合处理数值型数据类型</td>
</tr>
</tbody></table>
<hr>
###模型算法
基于模型的方法给每一个聚类假定一个模型，然后去寻找能够很好满足这个模型的数据集。这种算法

潜在假定是：目标数据集是由一系列的概率分布决定的。
<table>
<tbody><tr>
<td>AutoClass</td>
<td>是以概率混合模型为基础，利用属性的概率分布来描述聚类，该方法能够处理混合型的数据，要求各属性相互独立</td>
</tr>
<tr>
<td>自组织神经网络SOM</td>
<td>由外界输入不同的样本到人工的自组织映射网络中，一开始时，输入样本引起输出兴奋细胞的置各不相同，但自组织后会形成一些细胞群，它们分别代表了输入样本，反映了输入样本的特征</td>
</tr>
</tbody></table>
<hr>
###weka 中的聚类算法

EM，用户可指定需要产生多少聚类，否则所用的算法可通过交叉验证来决定，用户可指定循环数的最大值，并且为正常的密度计算设定可允许的最小标准差。

SimpleKMeans 使用 k 均值来聚类数据；聚类的数量通过一个参数设定。

Cobweb 实现了用于名词属性的 Cobweb 算法和用于数值性属性的 Classit 算法。

FarthestFirst 实现 Hochbaum 和 Shmoys 远端优先遍历算法。

MakeDensityBaseCluster 是一个元聚类器，它包装一个聚类算法，使其返回一个概率分布和密。它为每个聚类拟合一个离散分布，或一个对称的正态分布。

