


第12章 离群点检测

第12章：离群点检测

- 离群点分析 
- 基于统计学的方法
- 基于距离的方法
- 基于偏离的方法

离群点分析

- 什么是离群点？
 - 对象的集合, 它们与数据的其它部分不一致
 - 离群点可能是度量或执行错误所导致
 - 离群点也可能是固有的数据变异性的结果
- 问题：
 - 给定一个 n 个数据点或对象的集合, 及预期的离群点的数目 k , 发现与剩余的数据相比是相异的, 例外的, 或不一致的前 k 个对象
- 两个子问题：
 - 定义在给定的数据集合中什么样的数据可以被认为是不一致的
 - 找到一个有效的方法来挖掘这样的离群点

离群点分析

■ 应用：

- 信用卡欺诈检测
- 网络流量异常监测
- 顾客分割： 确定极低或极高收入的客户的消费行为
- 医疗分析： 发现对多种治疗方式的不寻常的反应

■ 离群点的类型

- 全局离群点： 数据对象显著的偏离数据集中的其余对象
- 情景离群点： 如果数据对象在给定特定情景下，显著的偏离其它对象
- 集体离群点： 数据对象的某个子集显著偏离整个数据集

离群点分析

- 采用数据可视化方法来进行离群点探测如何？
 - 不适用于包含周期性曲线的数据
 - 对于探测有很多分类属性的数据, 或高维数据中的离群点效率很低
- 方法
 - 统计学方法
 - 基于距离的方法
 - 基于偏差的方法
 - 基于密度的方法

基于统计学的离群点检测

- 对给定的数据集合假设了一个分布或概率模型(例如, 正态分布), 然后根据模型采用不一致性检验(**discordancy test**)来确定离群点
- 检验要求的参数
 - 数据集参数: 例如, 假设的数据分布
 - 分布参数: 例如平均值和方差
 - 和预期的离群点的数目

基于统计学的离群点检测

- 工作假设H是一个命题： n 个对象的整个数据集合来自一个初始的分布模型F
 - 即 $H: O_i \in F, i=1, 2, \dots, n$
- 不一致性检验：验证一个对象 O_i 关于分布F是否显著地大(或小)，即F产生 O_i 的概率是否足够小
- 主要的方法
 - 参数的方法：假设数据服从特定分布，分布的参数通过最大似然估计得到
 - 非参数方法：够造直方图，检测数据是否落入直方图的某一箱中

基于统计学的离群点检测

■ 缺点

- 绝大多数检验是针对单个属性的，而许多数据挖掘问题要求在多维空间中发现离群点
- 统计学方法要求关于数据集合参数的知识(如, 数据分布), 但是在许多情况下, 数据分布可能是未知的
- 当没有特定的检验时, 统计学方法不能确保所有的离群点被发现; 或者观察到的分布不能恰当地被任何标准的分布来模拟

基于距离的离群点检测

- 为了解决统计学方法带来的一些限制，引入了基于距离的离群点的概念
- 基于距离的离群点：
 - $DB(p, d)$ -离群点是数据集 T 中的一个对象 o ，使得 T 中的对象至少有 p 部分与 o 的距离大于 d
- 将基于距离的离群点看作是那些没有“足够多”邻居的对象（邻居是基于距给定对象的距离来定义的）
- 对许多不一致性检验来说，如果一个对象 o 根据给定的检验是一个离群点，那么对恰当定义的 p 和 d ， o 也是一个 $DB(p, d)$ 离群点

基于偏离的离群点检测

- 通过检查一组对象的主要特征来确定离群点，与给出的描述偏离的对象被认为是离群点
- 两个重要的概念：
 - 异常集(exception set): 它是偏离或离群点的集合, 被定义为某类对象的最小子集, 这些对象的去除会导致剩余集合的相异度的最大减少
 - 相异度函数(dissimilarity function): 是满足如下条件的任意函数: 当给定一组对象时, 如果对象间相似, 返回值就较小; 对象间的相异度越大, 函数返回的值就越大

例: 给定 n 个对象的子集合 $\{x_1, \dots, x_n\}$, 一个可能的相异度函数是集合中对象的方差

基于偏离的离群点检测

- 平滑因子(smoothing factor):
 - 一个为序列中的每个子集计算的函数.
 - 它估算从原始的数据集合中移走子集合可以带来的相异度的降低程度.
 - 平滑因子值最大的子集是异常集

一般的寻找全局最优的异常集的任务是NP难问题

基于偏离的离群点检测

- 一个顺序的方法在计算上是可行的, 能够用一个线性的算法实现
 - 不考虑估算当前子集关于其补集的相异度, 该算法从集合中选择一个子集合的序列来分析
 - 对每个子集合, 它确定其与序列中前一个子集合的相异度差异
 - 为了减轻输入顺序对结果的任何可能的影响, 以上的处理过程可以被重复若干次, 每一次采用子集合的一个不同的随机顺序
 - 在所有的迭代中有最大平滑因子值的子集合成为异常集

课程结束
感谢各位同学的支持！