第5章

关系数据可视化

本章的内容是关系数据在大数据中的应用及图形表示方法，主要介绍数据关联性的处理与数据分布性的处理。


第5章

5.1 关系数据在大数据中的应用

大数据的一个重要价值是可以帮助人们找到变量之间的联系，发掘事物背后的因果关系。在进行大数据挖掘前的重要一步就是探索变量的相关关系，进而才能探索背后可能隐藏着的因果关系。

分析数据时，不仅可以从整体进行观察，还可以关注数据的分布，如数据间是否存在重叠或者是否毫不相干？还可以从更宽泛的角度观察各个分布数据的相关关系。其实最重要的一点，就是数据在进行可视化处理后，呈现在读者眼前的图表所表达的意义是什么。

关系数据具有关联性和分布性。下面通过实例具体讲解关系数据，以及如何观察数据间的相关关系。

5.2 数据的关联性

数据的关联性，即数据相关性，是指数据之间存在某种关系。数据相关分析具有可以快捷、高效地发现事物间内在关联的优势，有效地应用于推荐系统、商业分析、公共管理、医疗诊断等领域。

事物之间的关联性是比较容易被发现的，但是关联并不代表存在因果关系。例如，大豆价格上涨，猪肉的价格可能也会上涨，但是大豆的价格上涨可能不是猪肉上涨的原因。

尽管如此,关联性还是能带来巨大的价值的,如大豆的价格已经上涨了,那我们就可以抓紧时间囤一些猪肉,这样往往能省下一笔钱,至于背后是否存在因果关系,就没那么重要了。大数据可视化就是在告诉我们分析结果是“什么”,而不是“为什么”.

数据的关联性,其核心就是指量化的两个数据间的数理关系。关联性强,是指当一个数值变化时,另一个数值也会随之相应地发生变化。相反地,关联性弱,就是指当一个数值变化时另一个数值几乎没有发生变化。通过数据关联性,就可以根据一个已知的数值变化来预测另一个数值的变化。下面通过散点图、散点图矩阵、气泡图来研究这类关系。

5.2.1 散点图

第 3 章已经介绍了以时间为横轴的散点图,这类散点图可以理解为用于发现数据和时间之间的关联关系。将横轴替换为其他变量,就可以用来比较跨类别的聚合数据。一般有三种关系:正相关、负相关和不相关,如图 5-1 所示。正相关时,横轴数据和纵轴数据变化趋势相同;负相关时,横轴数据和纵轴数据变化趋势相反;不相关时,散点的排列则是杂乱无章的。在统计学中有更科学的方法(如相关系数)衡量两个变量的相关性,但是散点图往往是判断相关性最简单、直观的方法,在计算相关系数前通常依靠散点图做出初步判断。

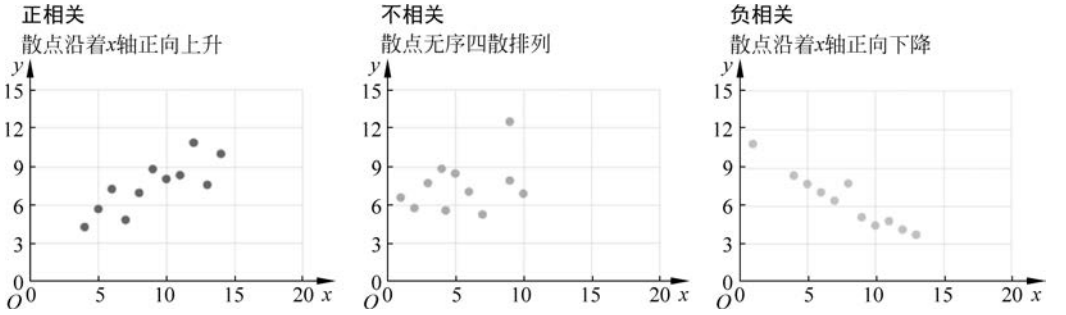

图 5-1 散点图与相关性判断示例

使用散点图时要注意以下几个问题。

当要在不考虑时间的情况下比较大量数据点时,常使用散点图。

(1) 即便自变量为连续性变量,仍然可以使用散点图。

(2) 如果在散点图中有多个序列,考虑将每个序列中点的标记形状更改为方形、三角形、菱形或其他形状,以示区别。

(3) 散点图中包含的数据越多,比较的效果就越好。

5.2.2 散点图矩阵

散点图矩阵是借助两变量散点图的作图方法,它可以看作是一个大的图形方阵,其每个非主对角元素的位置上是对应行的变量与对应列的变量的散点图,而主对角元素位置上是各变量名。

借助散点图矩阵可以清晰地看到所研究多个变量两两之间的相关关系,其基本框架如图 5-2 所示。

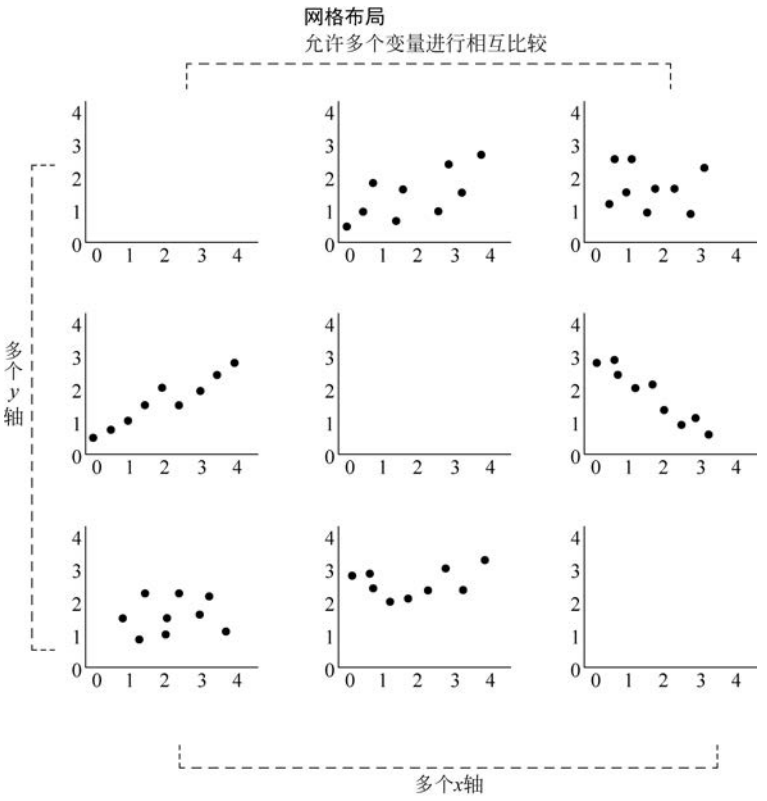

图 5-2　散点图矩阵

5.2.3　气泡图

气泡图和散点图相比，多了一个维度的数据。气泡图就是将散点图中没有大小的“点”变成有大小的“圆”，圆的大小用来表示多出的那一维数据的大小。气泡图让我们可以同时比较三个变量，其基本框架如图 5-3 所示。

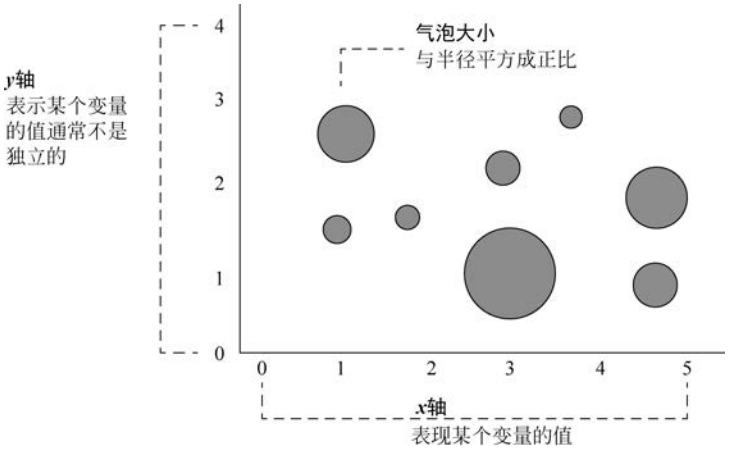

图 5-3　气泡图的基本框架

一个具体的例子如图 5-4 所示。二手车的价格由车龄和里程来决定,可以看出,两个指标越小,气泡越大,代表价格越高,反之则反。

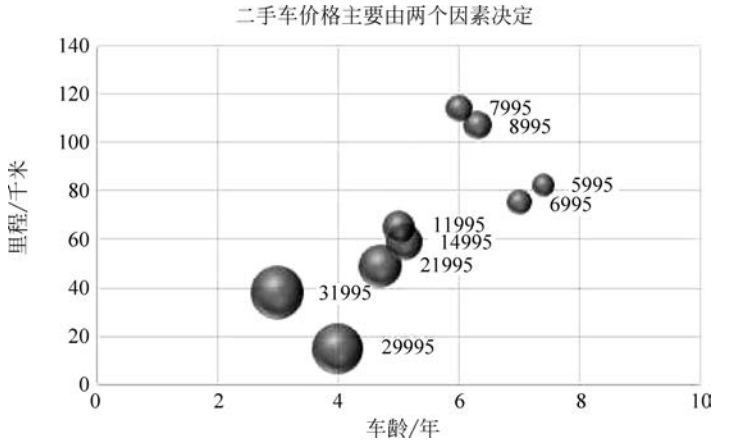

图 5-4 二手车车龄、里程与价格关系气泡图

5.3 数据的分布性

5.3.1 茎叶图

茎叶图又称“枝叶图”,是由 20 世纪早期的英国统计学家阿瑟·鲍利(Arthur Bowley)设计的。1997 年,统计学家约翰托奇(John Tukey)在其著作《探索性数据分析》(*Exploratory Data Analysis*)中将这种绘图方法介绍给大家,从此这种作图方法变得流行起来。茎叶图示例如图 5-5 所示。

茎	树叶
0	5 8
1	2 3 5 7
2	0 0 0 5 8 8 9
3	0 0 1 3 3 3 6 6 7 7 7 7 7 8 8 8 8 9 9
4	1 3 5 5 5 6 7 7 8 8 8 9 9
5	0 0 0 1 1 1 1 2 6 8
6	0 0 1 1 2 4 4 4 4 8 8 9
7	0 5 5 5 5 7
8	3 4 4 5 6 6 6 7 8 9
9	0 1 2 2 2 2 5 5 6 8 9 9
10	2 2 2 5 7

原始数据:
102, 102, 102, 105, 107

图 5-5 茎叶图示例

茎叶图的思路是将数组中的数按位数进行比较，将数的大小基本不变或变化不大的位作为一主干(茎)，将变化大的位的数作为分枝(叶)，列在主干的后面，这样就可以清楚地看到每个主干后面有几个数，每个数具体是多少。

茎叶图是一个与直方图相类似的特殊工具，但又与直方图不同，茎叶图保留原始资料的信息，直方图则失去原始资料的信息。将茎叶图茎和叶逆时针方向旋转 90°，实际上就是一个直方图，可以从中统计出次数，计算出各数据段的频率或百分比，从而看出分布是否与正态分布或单峰偏态分布逼近。

茎叶图的优点是统计图上没有原始数据信息的损失，所有数据信息都可以从茎叶图中得到。茎叶图中的数据还可以随时记录，随时添加，方便记录与表示。

茎叶图的缺点是只便于表示个位之前相差不大的数据，而且茎叶图只方便记录两组的数据。

5.3.2　直方图

直方图与茎叶图类似，若逆时针翻转茎叶图，则行就变成列；若是把每一列的数字改成柱形，则得到一个直方图。直方图又称质量分布图，是数值数据分布的精确图形表示。直方图中的柱形高度表示的是数值频率，柱形的宽度是取值区间。水平轴和垂直轴与一般的柱形图不同，它是连续的；一般的柱形图的水平轴是分离的，如图 5-6 所示。

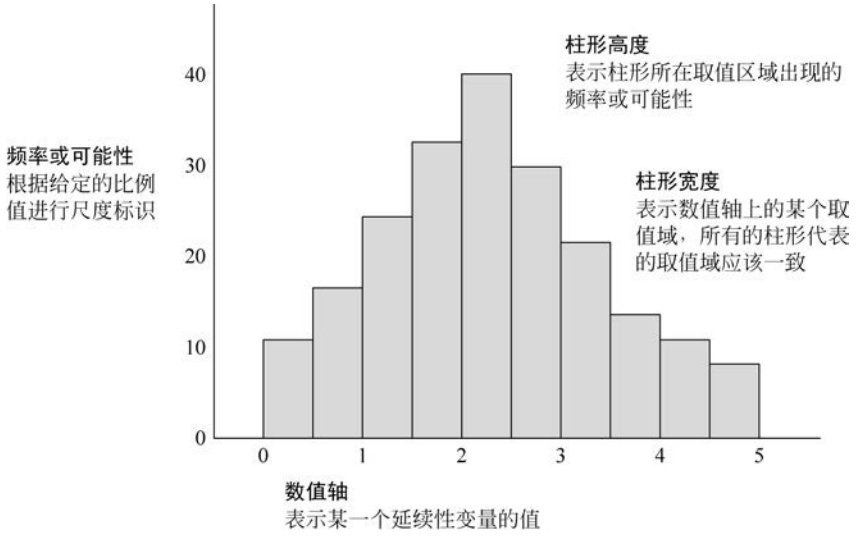

图 5-6　直方图的基本框架

5.3.3　密度图

直方图反映的是一组数据的分布情况，直方图的水平轴是连续性的，整个图表呈现的是柱形，用户无法获知每个柱形的内部变化。而在茎叶图中，用户可以看到具体数字，但是要求比较数值间的差距大小并不是很明确。为了呈现更多的细节，人们提出了密度图，可用它对分布的细节变化进行可视化处理。

当直方图分段放大时,分段之间的组距就会缩短,此时依照直方图画出的折线就会逐渐变成一条光滑的曲线,这条曲线就称为总体的密度分布曲线。这条曲线可以反映数据分布的密度情况,其基本框架如图 5-7 所示。

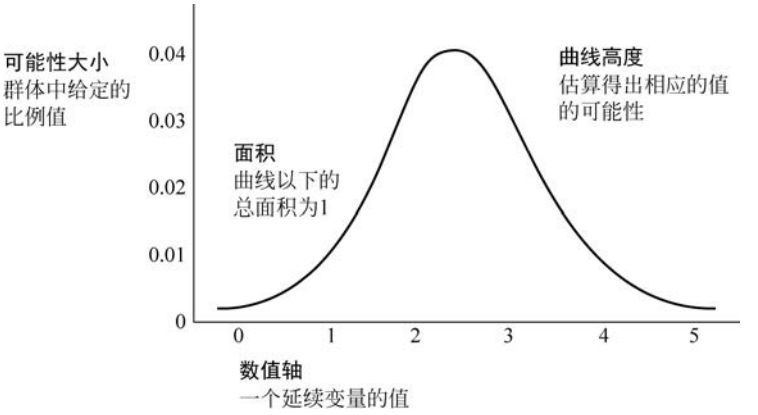

图 5-7 密度图的基本框架

习 题

1. 对于原始数据,如何初步判断关联性?
2. 直方图中面积有何意义?
3. 查询资料,找到一些常见的密度图,并解释它们的含义。