

第 5 章

大数据分析 with 大数据挖掘

第 4 章讲述了大数据采集与预处理,数据经过采集与预处理后,接下来要做的工作就是要对数据进行分析,数据分析可以分为广义的数据分析和狭义的数据分析、狭义的数据分析是我们常说的数据分析,广义的数据分析包括狭义的数据分析和数据挖掘。数据分析就是为了从数据中发现规律性的信息,帮助企业/个人预测未来的趋势和行为,做出具有针对的决策,从而使得商务和生产活动具有前瞻性。数据分析是数学与计算机科学相结合的产物,在实际应用中,人们可以通过计算工具和数学知识处理数据,得出结果,做出判断,以便采取适当行动。本章主要讲述数据分析、数据挖掘的方法,以及常用的工具,帮助人们从数据中发现规律,提供决策参考。

5.1 大数据分析的基本概念

大数据作为时下最火热的 IT 行业的词汇,随之而来的有数据仓库、数据安全、数据分析、数据挖掘等,围绕大数据的商业价值的利用逐渐成为行业人士争相追捧的利润焦点。随着大数据时代的来临,大数据分析应运而生。

5.1.1 数据分析概论

大数据分析是指用适当的统计方法对收集来的大量数据进行分析,为了提取有用信息和形成结论而对数据加以详细研究和概括总结的过程。在实际应用中,大数据分析可以帮助人们做出判断,以便采取合适的行动或措施。数据分析的数学基础在 20 世纪早期就已经确立,但直到计算机的出现才使实际操作成为可能,并使数据分析得以推广。数据分析是数学与计算机科学相结合的产物。

大数据分析的目的是把隐藏在一大批看来杂乱无章的数据中的信息集中和提炼出来,从而找出所研究对象的内在规律。在实际应用中,数据分析可以帮助人们做出判断,以便采取适当行动。数据分析是有组织、有目的地收集数据、分析数据,使之成为信息的过程。这一过程是质量管理体系的支持过程。在产品的整个生命周期,包括从市场调研到售后服务和最终处置的各个过程都需要适当运用数据分析过程,以提升有效性。

大数据分析是对规模巨大的数据进行分析,挖掘数据的有利信息并加以有效利用,将数据的深层价值体现出来。大数据分析之所以备受关注,本质原因是因为大数据具有巨大的潜在价值。大数据分析技术作为大数据获取数据价值的关键手段,在大数据分析中占有极其重要的位置,可以说是决定大数据价值能否发掘出来的关键因素。数据分析是整个大数据处理流程的核心。在数据分析的过程中,人们采用适当的方法(包括统计分析和数据挖掘

等方法),对采集到的海量数据进行详细研究和概括总结,从而发现和利用其中蕴含的信息和规律。大数据分析的主要目的是推测或解释数据、检查数据是否合法、给决策提供合理建议、诊断或推断错误原因以及预测未来将要发生的事情。正是有了大数据分析才能让规模巨大的数据有条有理、正确分类,产生有价值的分析报告,从而应用到各领域,促进其发展。

5.1.2 数据分析的类型

根据大数据的数据类型,可以把大数据分析划分成如下3类。

(1) 结构化数据分析:对传统关系数据库数据的分析。

结构化数据是可以以固定格式存储、访问和处理的任何数据。在一段时间内,计算机科学领域的人才在开发用于处理此类数据的技术方面取得了更大的成功(这种格式已经众所周知),并从中获得了价值。但是,如今,当此类数据的大小大幅增长时,人们可以预见的问题是,典型的数据大小正处于多个 ZB 中。关系数据库管理系统中存储的数据就是结构化数据的一个示例,如图 5-1 所示。

员工ID	员工姓名	性别	部	Salary_in_lacs
2365	拉杰什·库尔卡尼 (Rajesh Kulkarni)	男	金融	650000
3398	Pratibha Joshi	女	管理员	650000
7465	舒希尔·罗伊	男	管理员	500000
7500	Shubhojit Das	男	金融	500000

图 5-1 结构化数据示例

(2) 半结构化数据分析:对 HTML 网页或者 XML 文档等半结构化数据的分析。

半结构化数据可以包含 HTML 网页或者 XML 文档等两种形式的数据。可以将半结构化数据视为结构化的形式,但实际上并没有使用例如关系数据库管理系统(DBMS)中的表定义进行定义。半结构化数据的示例是 XML 文件中表示的数据,如图 5-2 所示。

```
<rec> <name>闷闷不乐</ name> <sex>男性</ sex> <age> 35 </ age> </ rec>
<rec> <name> Seema R.</ name> <sex>女性</ sex> <age> 41 </ age> </ rec>
<rec> <name> Satish Mane </ name> <sex> Male </ sex> <age> 29 </ age> </ rec>
<rec> <name> Subrato Roy </ name> <sex> Male </ sex> <age> 26 </ age> </ rec>
```

图 5-2 半结构化数据示例

(3) 非结构化数据分析:对图像、声音和视频等非结构化数据的分析。

任何形式或结构未知的数据都归为非结构化数据。除了庞大的数据量外,非结构化数据在处理中获得价值的过程中也带来了许多挑战。非结构化数据的典型示例是异构数据源,其中包含简单文本文件、图像、视频等的组合。如今,组织机构拥有大量可用数据,但不幸的是,他们不知道如何从中获取价值,此数据为原始格式或非结构化格式。

大数据分析的出现不是对传统数据分析的否定,而是对传统数据分析的集成和发展。

传统数据分析方法中的数据挖掘和统计分析仍然在大数据分析中发挥着重要作用。同时大数据分析也呈现出和传统数据分析不同的特征,表现在如下4方面。

① 分析的数据量不一样。传统的数据分析是对少量的数据样本进行分析,而大数据分析的是与事物相关的所有数据,而不是依靠分析少量的数据样本。

② 分析的侧重点不一样。大数据分析的重点不是发现事务之间的因果关系,而是发现事物之间的相关关系,因此相关分析是大数据分析的重要内容。

③ 分析的数据来源不一样。传统数据分析的对象大多局限在同一个来源的数据,但大数据分析更强调数据的融合,多种来源的原始数据进行融合才能反映事物的全貌。

④ 数据的解释方式不一样。可视化分析在传统数据分析中只是一种辅助分析手段,但大数据分析中更强调可视化分析的应用。

5.2 大数据分析方法

5.2.1 数据分析方法概述

尽管数据的目标和应用领域不同,一些常用的分析方法对大数据同样适用。

(1) 数据挖掘:数据挖掘是发现大数据集中数据模式的计算方法。很多数据挖掘算法已经在人工智能、机器学习、模式识别、统计和数据库领域得到了广泛应用。数据挖掘主要采用决策树、神经网络、关联规则、聚类分析等统计学、人工智能、机器学习等方法进行挖掘;著名的数据挖掘算法包括决策树、k-means 算法、支持向量机、Apriori 算法、PageRank 算法、朴素贝叶斯分类等,覆盖了分类、聚类、回归和统计学习等方面,5.3 节重点讲述数据挖掘相关知识。

(2) 数据分析:在统计理论中,通过概率理论对数据的随机性和不确定性建立模型。统计分析技术可以分析描述性统计分析和推断性统计技术。描述性统计技术对数据集进行总结或描述,而推断性统计技术则能够对过程进行推断。统计分析方法主要有对比分析、回归、因子分析、聚类等。数据分析一般都是得到一个指标统计量结果,如总和、平均值等,这些指标数据都需要与业务结合进行解读,才能发挥出数据的价值与作用。本节的数据分析重点讲述数据分析相关知识。下面介绍几种数据分析方法。

1. 相关分析

事物之间往往存在某种关联性,如果这种关联性可以用函数表示,则称它们之间是一种函数关系。现实中很多事物之间虽然存在某种联系,但不能应用已知的函数关系来表示,这种联系即为相关关系。如果这种相关性只涉及两个事物则为单相关,如果涉及3个或者3个以上的事物则为复相关、多重相关。

事物之间的相关程度使用相关系数来衡量,相关系数表示事物之间关系的紧密程度。对于复相关,往往采用多重相关系数考察一个变量与其他变量之间的相关程度,采用偏相关系数考察多个变量中两个变量的相关性。

在有 n 个($n \geq 3$)变量的系统中,若要考察第 i 个变量与其余 $n-1$ 个变量的相关程度,采用多重相关系数来表示,计算公式为 $\sqrt{1 - \frac{R}{R_{ii}}}$ 。 R 是单相关系数矩阵对应的行列式, R_{ii}

是 R 的第 i 行、第 i 列的代数余子式。 R_{ii} 的代数余子式是在去掉 R 中第 i 行与第 i 列元素得到的行列式,同理 R_{ij} 的代数余子式是在去掉 R 中第 i 行与第 j 列元素得到的行列式。

多重相关性中考察一个变量与另外一个变量之间的相关性用偏相关系数来表示。例如考察变量 i 与变量 j 之间的偏相关性,计算公式为 $(-1)^{(i+j)} \frac{R_{ij}}{\sqrt{R_{jj}R_{ii}}}$ 。该值的绝对值越大,表明变量 i 和 j 的偏相关程度越大,二者的关系越紧密,相互影响越明显。

2. 回归分析

回归分析是一种统计分析方法,用于确定变量之间的函数关系,主要用于函数的预测。回归分析的思想是根据若干变量的一系列的实际观测值,推断出这些变量之间存在的函数关系,然后再利用所获得的函数关系预测某个变量的取值。如果回归分析只涉及两个变量且二者的关系可以表示为线性函数时则称为一元线性回归分析;如果回归分析中包含 3 个或 3 个以上的变量且变量之间可以表示成线性函数则称为多重线性回归分析。

进行回归分析时,需要使用残差来衡量回归分析结果的优劣。残差是预测值和实际观测值之间的差额。当获得一个回归分析的函数关系时,对于给定的自变量,可以算出因变量的值,但是这种函数只是尝试去逼近真实的情况,由于随机误差等因素,根据函数关系计算得到的因变量的值(又称预测值)与实际观测值有一定的差距。残差就是用来衡量其大小的指标。残差越小,说明预测值和实际观测值越接近,回归分析的结果也越好。

3. 聚类分析

聚类分析属于探索性的数据分析方法。通常,利用聚类分析可以将看似无序的对象进行分组、归类,以达到更好地理解研究对象的目的。聚类结果要求组内对象相似性较高,组间对象相似性较低。在用户研究中,很多问题都可以借助聚类分析来解决,如网站的信息分类问题、网页的点击行为关联性问题以及用户分类问题等。其中,用户分类是最常见的情况。

聚类分析是根据数据的数值特征对数据进行分类的一种分析方法。与一般的分类算法不同,聚类分析并不能确定数据应该分为几类。聚类分析的目的是将众多的个体先聚集成比较好处理的几个类别或子集,然后再利用判别分析进一步研究各个类别之间的常见情况。

对一组数据,既可以对变量(指标)进行聚类分析,也可以对观测值进行聚类分析。分析的时候,不一定要事先假定有多少类,也可以完全根据数据自身的规律来分类。一般将变量的聚类分析称为 R 型聚类,而将观测值聚类称为 Q 型聚类。

聚类分析中,比较重要的是两个距离的概念,按照远近程度来聚类是聚类分析法的要义,那么这个远近究竟指什么呢?这里的距离一方面是指点与点之间的距离,另一方面是指类和类之间的距离。点间距离本身有多个定义方式也即多种运算方法,因此只要选择一种算法即可。由一个点组成的类是最基本的类,如果每一类都由一个点组成,那么点间距离就是类间距离。但如果一个类包含不止一个点,那么就需要确定类间距离。类间距离是基于点间距离定义的,如两类之间最近点之间的距离可以作为两类间距离,也可以选用最远点的距离,还可以选择各类之间的中心距离。

聚类分析有多种方法,不同的系统提供了不同的聚类分析法。SPSS 提供了 K-平均值

聚类、两步聚类和系统聚类 3 种聚类方法,但它们的应用范围和优劣势各有不同。

K-平均值聚类(KCA)又称为大快速聚类,是进行人群细分时最常使用的方法。该方法是单纯应用统计技术根据若干指定变量(应限制为尺度变量)将众多个案分到固定的类别中去。这种方法用于大量(数千)个案的类别划分时非常有效。但该方法可以选择的内容较少,最重要的是选择聚类的数量、迭代的次数以及聚类的中心位置,所以人为经验和判断无形中会起很大作用。KCA 方法本身不仅要求确定分类的类数,而且需要事先确定点,也就是类种子。在实际操作中,SPSS 会自动选取种子,然后根据其他点离这些种子的远近对所有点进行分类。再然后,就是将这几类的中心(均值)作为新的基石再分类,如此迭代。

两步聚类是揭示自然类别的探索性工具。该方法的算法与传统聚类技术相比有一些显著的特点:它可以基于类别变量和连续变量来进行聚类,自动选择聚类结果的最佳类别数;具备有效分析大量数据的能力。

如果只拥有少量的个案(少于数百个),并且想尝试多种聚类方法,测量不同类别之间的差异,则应该尝试使用系统聚类。系统聚类也叫作层次聚类(HCA)。当然,该方法不仅可以对样本聚类,还可以对变量聚类。这种方法的分类结果取决于对聚类方法、距离测量方法、标准化变量的设置。这种方法不事先确定类数,有多少点就是多少类,它沿着最近的距离先聚为一类的思想进行合并,直至最后只有一个大类为止。

5.2.2 数据分析过程

数据分析,是有目的地进行收集、整理、加工和分析数据,提炼有价值信息的过程,其过程概括起来主要包括:明确分析目的与思路、数据收集、数据处理、数据分析、数据展现和撰写报告 6 个阶段,如图 5-3 所示。

1. 明确数据分析的目的和思路

一个分析项目,数据对象是谁?目的是什么?要解决什么业务问题?数据分析师对这些都要了然于心。数据分析师首先要做的就是整理分析框架和分析思路。不同的项目对数据的要求,使用的分析手段也是不一样的。

2. 数据收集

数据收集是按照确定的数据分析和框架内容,有目的地收集、整合相关数据的过程,它是数据分析的基础。

3. 数据处理

数据处理是指对收集到的数据进行加工、整理,以便开展数据分析,它是数据分析前必不可少的阶段。这个过程是数据分析整个过程中最花费时间的,也在一定程度上取决于数据仓库的搭建和数据质量的保证。数据处理主要包括数据清洗、数据转化等。

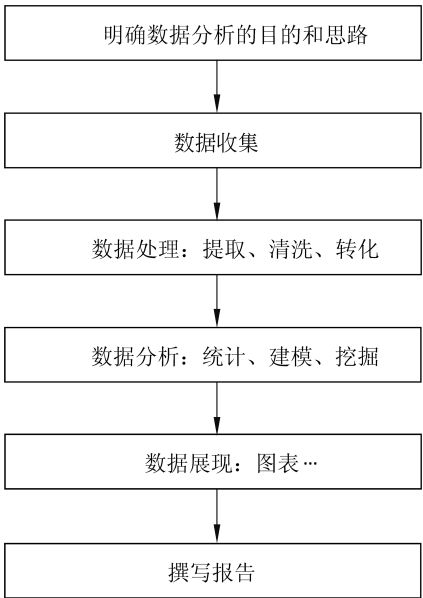


图 5-3 大数据分析过程

4. 数据分析

数据分析是指通过分析手段、方法和技巧对准备好的数据进行探索、分析,从中发现因果关系、内部联系和业务规律,并提供决策参考。

到了这个阶段,要能驾驭数据、开展数据分析,就要设计工具和方法的使用,其一要熟悉常规数据分析方法,最基本的是要了解例如方差、回归、因子、聚类、分类、时间序列等多元和数据分析方法的原理、使用范围、优缺点和结果的解释;其二是熟悉数据分析工具,Excel、SPSS、SAS 等是最常见的,一般的数据分析可以通过 Excel 完成,而后要熟悉至少一个专业的分析软件。

5. 数据展现

一般情况下,数据分析的结果都是通过图、表的方式来呈现的。借助数据展现手段,能更直观地让数据分析表述想要呈现的信息、观点和建议。

常用的图表包括饼图、折线图、柱形图/条形图、散点图、雷达图、金字塔图、矩阵图、漏斗图、帕累托图等。

6. 撰写报告

最后阶段就是撰写数据分析报告,这是对整个数据分析成果的呈现。通过分析报告,把数据分析的目的、过程、结果及方案完整呈现出来,提供决策参考。

一份好的数据分析报告,首先需要有好的分析框架,并且图文并茂、层次清晰,能够让阅读者一目了然。结构清晰、主次分明可以使阅读者正确理解报告内容;图文并茂可以令数据更加生动活泼,提高视觉冲击力,有助于阅读者更形象、直观地看清楚问题和结论,从而产生思考。

另外,数据分析报告需要有明确的结论、建议和解决方案,不仅是找出问题,后者更为重要,否则就不是好的分析,同时也失去了报告的意义。数据的初衷就是为解决一个目的才进行的分析,不能舍本求末。

5.2.3 数据处理结果分析

1. 相关分析实例

某上市公司 8 月前 15 个交易日的股票价格、成交金额和收益率的样本数据如表 5-1 所示。现要计算:

表 5-1 8 月份股票交易样本数据

日 期	股票价格/元	成交金额/元	收 益 率
20080801	9.28	41 652 766	0.014 923
20080802	9.23	18 716 130	-0.0023
20080803	9.18	41 314 097	-0.002 315
20080804	8.96	18 393 783	0.003 234

续表

日 期	股票价格/元	成交金额/元	收 益 率
20080805	8.95	34 259 522	—0.007 53
20080806	8.73	31 981 311	—0.025 97
20080807	8.65	43 000 708	—0.011 11
20080808	8.59	35 314 780	0.011 236
20080809	8.52	34 774 469	0.039 535
20080810	8.49	3 288 839	0.002 237
20080811	8.42	23 306 213	—0.018 97
20080812	8.37	38 787 086	—0.014 79
20080813	8.31	30 253 320	0.020 785
20080814	8.26	41 662 276	—0.002 26
20080815	8.21	23 703 595	0.002 245

收益率与股票价格和成交金额的多重相关系数；
收益率与股票价格的偏相关系数；
收益率与成交金额的偏相关系数。
在 Excel 中,对表 5-1 中的数据进行如下分析。
第 1 步,将表 5-1 所示的数据录入 Excel 文件中,该文件中的数据样式如图 5-4 所示。

股票价格/元	成交金额/元	收益率
9.28	41652766	0.014923
9.23	18716130	-0.0023
9.18	41314097	-0.002315
8.96	18393783	0.003234
8.95	34259522	-0.00753
8.73	31981311	-0.02597
8.65	43000708	-0.01111
8.59	35314780	0.011236
8.52	34774469	0.039535
8.49	3288839	0.002237
8.42	23306213	-0.01897
8.37	38787086	-0.01479
8.31	30253320	0.020785
8.26	41662276	-0.00226
8.21	23703595	0.002245

图 5-4 股票交易数据

第 2 步,在该文件中单击“数据”功能区最右边的“数据分析”图标。在弹出的“数据分析”窗口中选择“相关系数”(见图 5-5),单击“确定”按钮,弹出如图 5-6 所示的“相关系数”窗口。

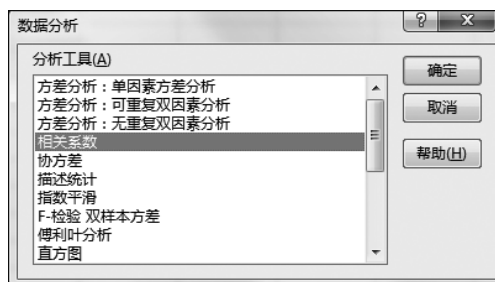


图 5-5 选择“相关系数”

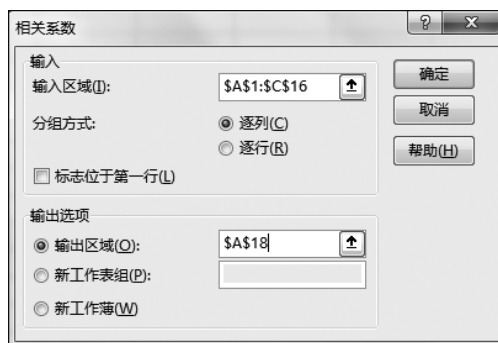


图 5-6 “相关系数”窗口

第 3 步,在如图 5-6 所示的“相关系数”窗口中,在“输入区域”选中\$A\$1:\$C\$16 区域;在“分组方式”中选中“逐列”,勾选“标志位于第一行”;在“输出选项”部分选中“输出区域”,并在后边的文本框中选中\$A\$18。

第 4 步,单击“确定”按钮,从单元格 A18 开始将显示如图 5-7 所示的矩阵信息。

18		股票价格 (元)	成交金额 (元)	收益率
19	股票价格 (元)	1		
20	成交金额/ (元)	0.085996155	1	
21	收益率	-0.018816503	0.020194853	1

图 5-7 单相关系数矩阵

第 5 步,根据对称性填充图中矩阵上方的空单元格,结果如图 5-8 所示。

18		股票价格 (元)	成交金额 (元)	收益率
19	股票价格 (元)	1	0.085996155	-0.0188165
20	成交金额/ (元)	0.085996155	1	0.02019485
21	收益率	-0.018816503	0.020194853	1

图 5-8 矩阵填充结果

第 6 步,列出 R 的行列式以及 R_{11} 、 R_{22} 、 R_{33} 、 R_{13} 、 R_{13} 的代数余子式,如图 5-9 所示。

第 7 步,计算的行列式的值。

选中单元格 B49,然后在公式栏中输入如下公式,如图 5-10 所示: =MDETERM (A25: C27)。

输完代码后按 Enter 键,计算结果如图 5-11 所示。

第 8 步,计算 R_{11} 的行列式的值。

选中单元格 B50,在公式栏中输入公式 =MDETERM(A30: B31)并按 Enter 键,计算结果如图 5-11 所示。

第 9 步,计算 R_{22} 的行列式的值。

选中单元格 B51,并在公式栏中输入公式 =MDETERM(A34: B35),并按 Enter 键,计算结果如图 5-11 所示,

第 10 步,计算 R_{33} 的行列式的值。

选中单元格 B52,在公式栏中输入公式 =MDETERM(A38: B39),并按 Enter 键,计算

	A	B	C	D
24	单相关系数矩阵			
25	1	0.085996155	-0.018816503	
26	0.085996155	1	0.020194853	
27	-0.018816503	0.020194853	1	
28				
29	R ₁₁			
30	1	0.020194853		
31	0.020194853	1		
32				
33	R ₂₂			
34	1	-0.018816503		
35	-0.018816503	1		
36				
37	R ₃₃			
38	1	0.085996155		
39	0.085996155	1		
40				
41	R ₁₃			
42	0.085996155	1		
43	-0.018816503	0.020194853		
44				
45	R ₂₃			
46	1	0.085996155		
47	-0.018816503	0.020194853		

图 5-9 各个矩阵数据

B49	:	  	=MDETERM(A25:C27)
-----	---	---	-------------------

图 5-10 计算 R 的公式

	A	B
48		
49	R	0.991777412
50	R ₁₁	0.999592168
51	R ₂₂	0.999645939
52	R ₃₃	0.992604661
53	R ₁₃	0.020553182
54	R ₂₃	0.021813

图 5-11 R、R11、R22、R33、R13、R23 的值

结果如图 5-11 所示。

第 11 步,计算 R_{13} 的行列式的值。

选中单元格 B53,在公式栏中输入公式=MDETERM(A42: B43),并按 Enter 键,计算结果如图 5-11 所示。

第 12 步,计算 R_{23} 的行列式的值。

选中单元格 B54,在公式栏中输入公式=MDETERM(A46: B47),并按 Enter 键,计算结果如图 5-11 所示。

第 13 步,计算收益率与股票价格和成交金额的多重相关系数: $\sqrt{1-\frac{R}{R_{33}}}$ 。

选中单元格 B56,在公式栏中输入公式=SQRT(1-B49/B52),并按 Enter 键,计算结果如图 5-12 所示。

	A	B
55		
56	多重相关系数	0.028868888
57	偏相关系数1	0.020633813
58	偏相关系数2	0.021897985

图 5-12 多重相关和偏相关系数值

第 14 步,计算收益率与股票价格的偏相关系数 1:

$$(-1)^{(1+3)} \frac{R_{13}}{\sqrt{R_{11}R_{33}}}。$$

选中单元格 B57,在公式栏中输入公式 =
(-1)⁽¹⁺³⁾ * B53/SQRT(B50 * B52),并按 Enter 键,

计算结果如图 5-12 所示。

第 15 步,计算收益率与成交金额的偏相关系数 2: $(-1)^{(1+3)} \frac{R_{23}}{\sqrt{R_{22}R_{33}}}$ 。

选中单元格 B58,在公式栏中输入公式 = (-1)⁽¹⁺³⁾ * B54/SQRT(B51 * B52),并按 Enter 键,计算结果如图 5-12 所示。

结果分析与总结:

图 5-12 反映的是收益率、股票价格和成交金额三者之间的相关性。单元格 B56 中反映的是收益率与股票价格和成交金额的多重相关系数,该值越大,表明收益率与股票价格和成交金额的线性相关程度越密切。单元格 B57 与 B58 分别是收益率与股票价格、收益率与成交金额的偏相关系数值。偏相关系数用于多要素组成的系统中,单独考察一个要素对其他要素的影响。其值取值范围介于-1 和 1 之间,绝对值越大表明其偏相关的程度越大。本实例中,相较于成交金额,股票价格对收益率的影响更大。

2. 回归分析实例

某媒体公司的管理者认为公司每周的收入与广告费用是密切相关的,他们想对每周的总收入做出预测和评估。这家公司收集了 8 周的历史数据组成样本数据,如表 5-2 所示。

表 5-2 收入与电视广告费、报纸广告费关系数据 (单位:千元)

每周的总收入	电视广告费用	报纸广告费用
96	5.0	1.5
90	2.0	2.0
95	4.0	1.5
92	2.5	2.5
95	3.0	3.5
94	3.5	2.3
94	2.5	4.2
94	3.0	2.5

现要进行如下两项工作: