Alibaba Group 阿里巴巴集团 | 技术丛书

Broadview www.broadview.com.cn

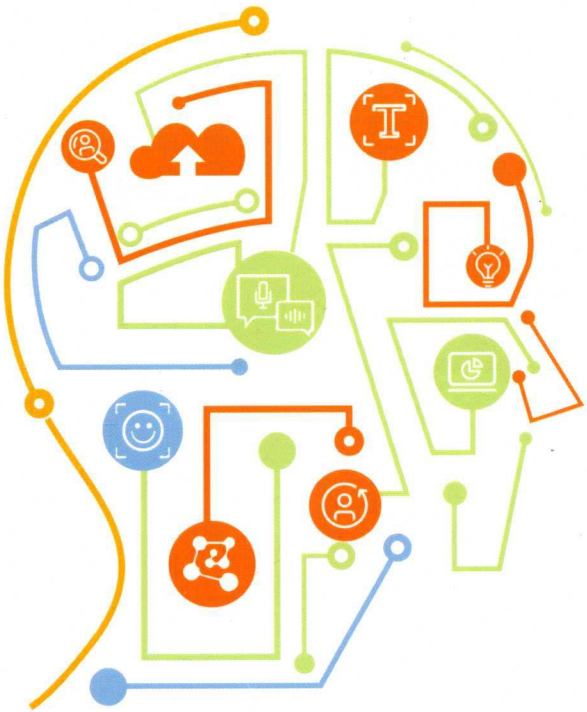
大数据丛书 | “十三五”国家重点图书出版规划项目

永远在线 | 实时响应 | 轻松上手 | 畅学无忧

机器学习在线

解析阿里云机器学习平台

杨旭 著



中国工信出版集团

電子工業出版社 PUBLISHING HOUSE OF ELECTRONICS INDUSTRY http://www.phei.com.cn

Alibaba Group 阿里巴巴集团 | 技术丛书

大数据丛书 | “十三五”国家重点图书出版规划项目

机器学习在线

解析阿里云机器学习平台

杨旭 著

電子工業出版社
Publishing House of Electronics Industry
北京•BEIJING

内 容 简 介

近几年，机器学习平台获得了飞速发展，积累了大量高效的机器学习算法组件，基于这些组件可以快速实现业务流程、解决具体问题。阿里云机器学习平台的丰富算法功能可以在线使用，不需要购买硬件，不需要安装配置各种环境；数据和计算资源一直处在“在线”状态，不必担心数据太大或计算资源不足的问题。机器学习平台降低了我们使用机器学习知识的门槛，将各个算法作为组件，即使不了解背后的理论知识，仍可以仿照书中实例，将组件连接起来解决一些实际问题。

本书适合机器学习算法的初学者及中级用户快速入门，在机器学习实践中学习。

图书在版编目（CIP）数据

机器学习在线：解析阿里云机器学习平台 / 杨旭著. —北京：电子工业出版社，2017.8
（阿里巴巴集团技术丛书）
ISBN 978-7-121-31869-6

Ⅰ. ①机… Ⅱ. ①杨… Ⅲ. ①电子商务－计算机网络 Ⅳ. ①TP393

中国版本图书馆 CIP 数据核字(2017)第 131004 号

策划编辑：刘　皎
责任编辑：郑柳洁
印　　刷：三河市鑫金马印装有限公司
装　　订：三河市鑫金马印装有限公司
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　邮编：100036
开　　本：787×980　1/16　印张：16.5　字数：263 千字
版　　次：2017 年 8 月第 1 版
印　　次：2017 年 8 月第 1 次印刷
定　　价：79.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888，88258888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

本书咨询联系方式：010-51260888-819，faq@phei.com.cn。

前言

飞速发展的互联网、物联网每时每刻都在产生大量的数据，数据的价值也因此被提升到前所未有的高度：越来越多的人投身数据分析的领域，希望通过机器学习及深度学习，从数据中获取更大的价值。另一方面，云计算的蓬勃发展极大地扩展了数据的存储能力，它使计算可以同时使用成百上千台机器，快速解决问题，而在计算完成后，又能及时释放掉资源，控制成本。

在这样的大背景下，机器学习算法平台也获得了飞速发展，积累了大量高效的机器学习算法组件，基于这些组件我们可以快速实现业务流程，解决具体问题。在为本书定书名时，受到王坚博士《在线》一书的影响，觉得用“在线”一词来说明目前机器学习平台的状态非常恰当：丰富的算法功能可以在线使用、不需要购买硬件、不需要安装配置各种环境；数据和计算资源一直处在“在线”状态，不必担心数据太大或计算资源不足的问题。

阿里云机器学习算法平台不仅在阿里集团内部使用，也已对阿里集团外部开放，读者可以通过阿里云官网试用或使用本书中介绍的功能。

机器学习平台提供了一个舞台，主角是其上面的近百种算法。本书的重点放在这些算法的使用上——通过实际的数据和具体的场景，帮助读者理解各算法所擅长处理的问题；另外，本书是根据机器学习的知识点由浅入深来逐步组织的，以降低阅读本书的门槛，使读者对所学的内容能产生清晰的印象。

在具体章节的组织上，阿里云机器学习平台的介绍占两个章节，即第 1 章和附录 A。第 1 章为平台简介，在内容组织上尽量减少文字说明，将最基本的内容用图例来表示；附录 A 介绍了些琐碎但重要的事情，像如何试用、如何上传数据以及预处理函数的详细说明。第 2 章至第 12 章是按照机器学习的知识点逐步深入的思路来编排的。分类模型是机器学习理论和应用方面

的重头，首先是数值类型特征的二分类模型、扩展特征的类型、多分类模型；之后介绍聚类模型；然后是回归模型；再后面介绍文本分析领域的应用（主题模型、向量化、关键词等），根据文本描述进行预测、情感分析，并以电影数据为例，搭建推荐系统。深度学习的内容放在第 12 章，围绕 TensorFlow 框架组件，介绍了一个能体现 TensorFlow 特点的 Softmax 模型的例子，然后介绍了使用深度学习 DNN 分类器的例子。

机器学习平台降低了我们使用机器学习知识的门槛，将各个算法作为组件，即使不了解其背后的理论知识，读者仍然可以仿照书中实例，将组件连接起来解决一些实际问题。希望本书能帮助读者在机器学习的实践中学习。

最后，感谢一起研发阿里云机器学习平台的各位同事！感谢家人的理解和支持！

杨旭

2017 年 7 月

目录

第1章 阿里云机器学习

阿里云机器学习平台是构建在阿里云 MaxCompute 计算平台之上，集数据处理、建模、离线预测、在线预测为一体的机器学习算法平台。用户通过拖曳可视化的操作组件来进行试验，使得没有机器学习背景的工程师也可以轻易上手玩转数据挖掘。平台提供了丰富的组件，包括数据预处理、特征工程、算法组件、预测与评估。平台目前整合了阿里集团内最先进的算法，为集团内、外不同用户提供算法服务。

欢迎访问阿里云机器学习的网址：https://data.aliyun.com/product/learn。用户可在阿里云网站申请公测，进行免费试用，相关内容详见本书附录。

1.1 产品特点

- 简单、易用

  将各个复杂的机器学习算法抽象为算法组件，通过拖曳组件的方式即可完成机器学习流程的搭建，大大降低了机器学习算法学习和使用的门槛。

- 算法丰富、完整

  不但包括了机器学习核心的分类、聚类、回归模型，还包括了数据探索、预处理、特征工程、深度学习、文本分析等方面的组件，可以一站式地完成不同场景的解决方案。

- 支持处理大数据

  提供高性能的机器学习算法实现，并根据数据量的大小及计算的复杂程度自动获取适合的计算资源，再多的数据也能及时处理。

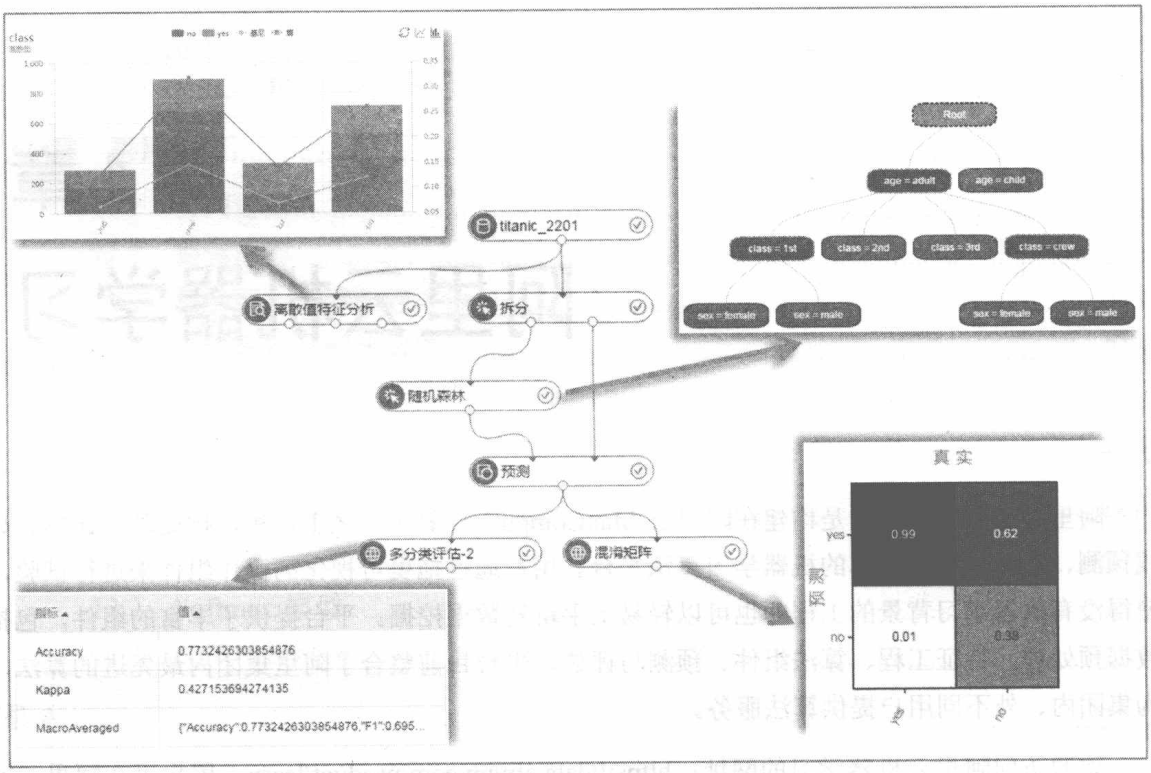


图 1-1　阿里云机器学习

1.2　名词解释

为便于读者阅读，将阿里云机器学习平台中涉及的一些名词进行了解释，详见表 1-1。

表 1-1　名词解释

名　词	解　释
MaxCompute	开放数据处理服务，由阿里云自主研发，提供针对 TB/PB 级数据、实时性要求不高的分布式处理能力，应用于数据分析、挖掘、商业智能等领域
项目（Project）	项目（也称项目空间）是 MaxCompute 最基本的组织对象。其他对象，例如表（Table）和实例（Instance）等都归属于某个项目
实验（Experiment）	实验是指阿里云机器学习平台用户搭建的数据工作流程或者数据应用。用户需要先建立一个实验实例，然后在实验画布上搭建数据流程

续表

名　词	解　释
MaxCompute 源表与 MaxCompute 目标表（Table）	表（Table）是 MaxCompute 中数据存储对象。与常见的关系型数据类似，MaxCompute 中的表逻辑上也是二维结构。源表指一个算法节点的输入，目标表指算法节点的输出
组件（Nodes）	组件是用户可以在阿里云机器学习平台上调用执行的最小操作单元，例如数据导入导出、数据处理、数据分析、模型训练或者预测
模型（Model）	模型是特指一个算法或者机器学习训练组件产生的结果数据。模型是一类特殊的组件
分区（Partition）	MaxCompute 表分区

1.3　构建机器学习实验

1.3.1　新建实验

如图 1-2 所示，点击左侧“实验”按钮，右击“我的实验”选项，选择“新建空白实验”或“从模板新建实验”选项，然后系统会自动进入新建的实验操作空间。

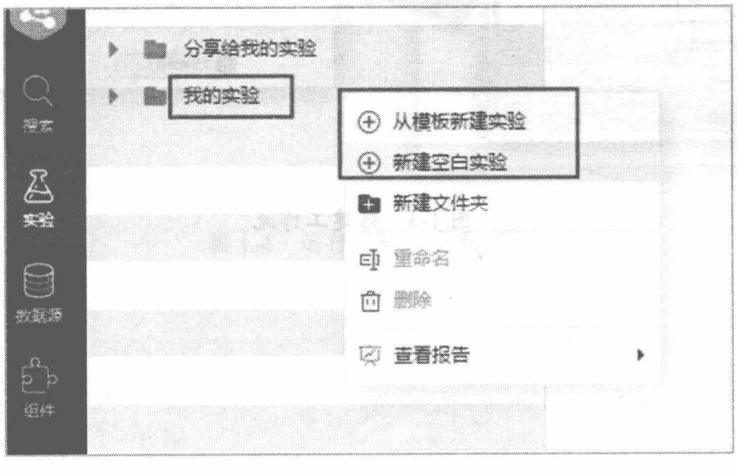


图 1-2　新建实验

1.3.2 使用组件搭建工作流

如图 1-3 所示，拉入数据表和算法组件，进行实验流搭建。具体操作是：点击左侧“数据源”按钮搜索选择需要的数据表，拖曳到右侧空白处；点击左侧“组件”按钮，选择需要的组件，并拖曳到右侧空白处；并根据实验流程，连接组件的输入、输出桩。

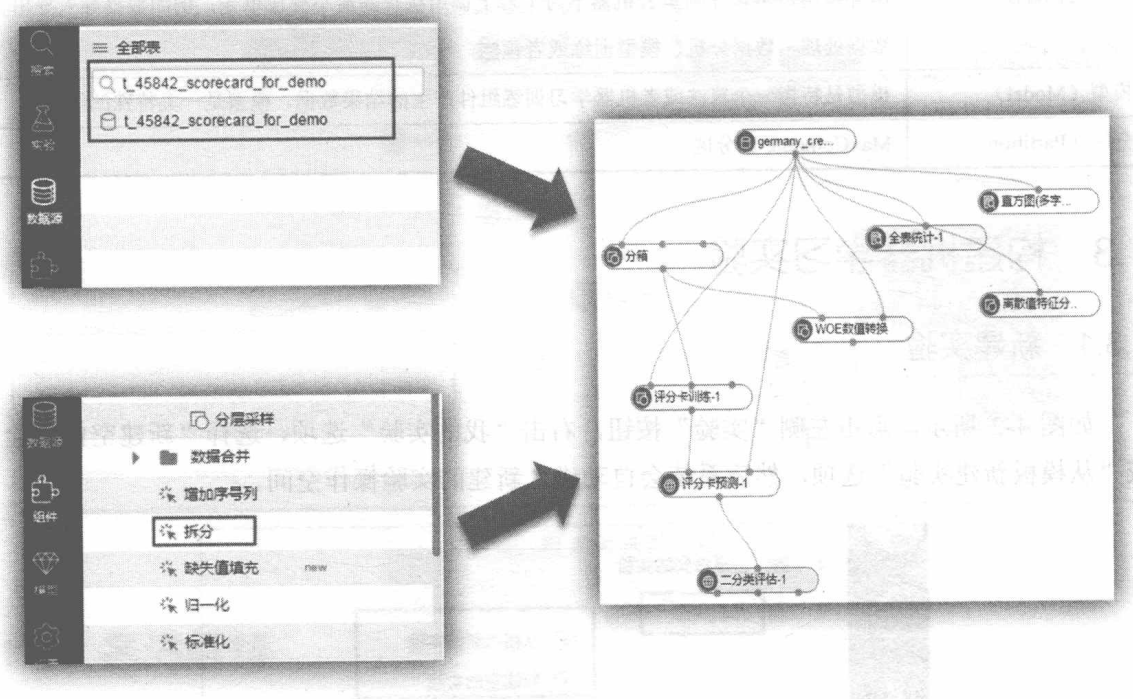


图 1-3 搭建工作流

1.3.3　运行实验、查看结果

如图 1-4 所示，点击工作区下方的“运行”按钮，依次运行实验的各个组件，组件运行完成后，其右端会显示绿色的对号标记，然后，单击鼠标右键，就可选择查看结果数据及图表。

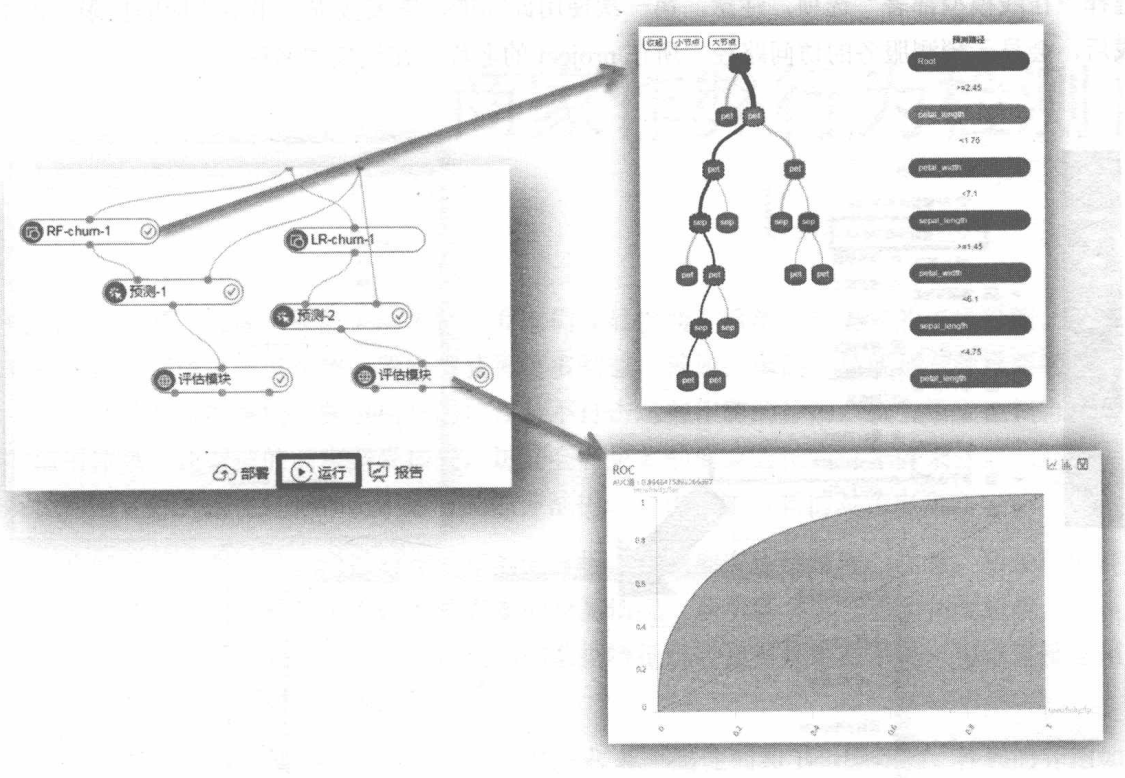


图 1-4　运行实验、查看结果

1.3.4 模型部署、在线预测

1）模型部署。

如图 1-5 所示，点击左侧“模型”按钮，找到当前实验名称，选择模型，然后在右键菜单选择“在线模型部署”选项。注意，第一次使用此功能，需要按提示申请相应的权限。部署完成后，会显示预测服务的访问路径、所在 project 的名称、在线模型名称。

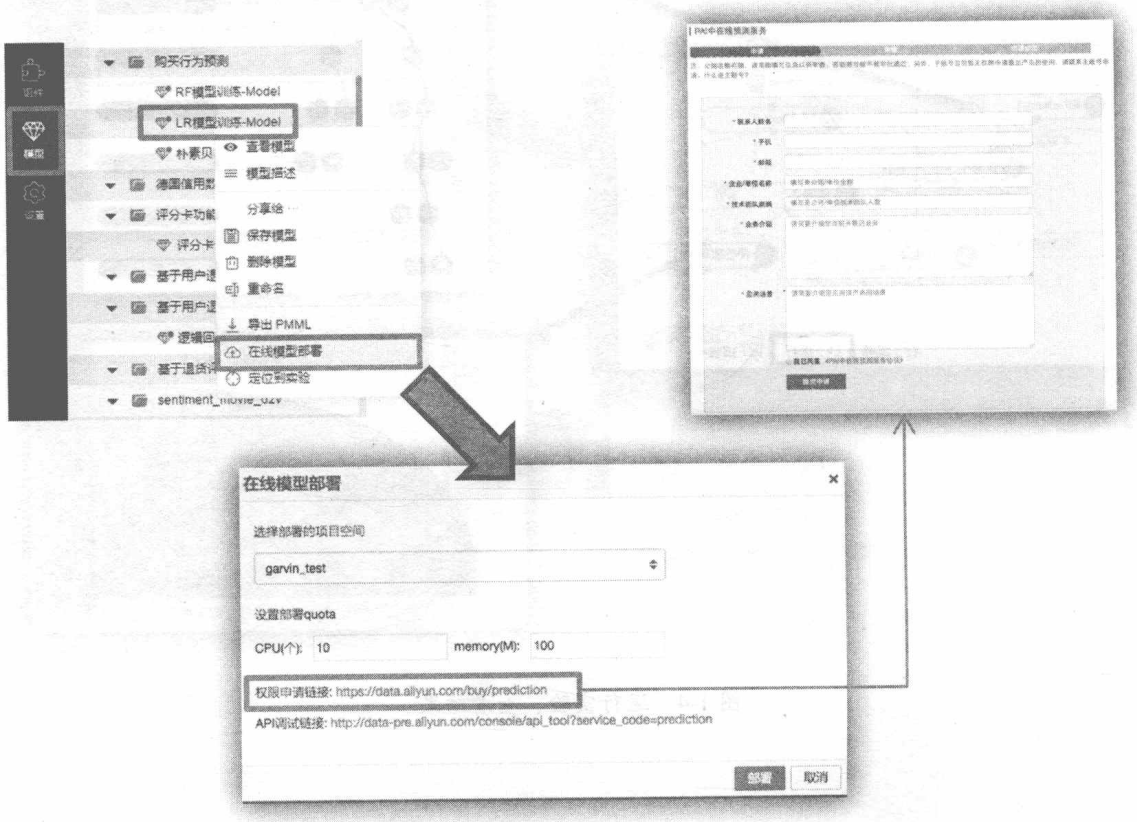


图 1-5 在线模型部署

2）在线预测。

预测 API 请求地址：$访问路径/projects/$project 名称/onlinemodels/$模型名称。

请求 Body 内容需要填上对应算法所输入的 json 格式文件，包括字段名、字段的 type 和具体数值。

第2章 商家作弊行为检测

电子商务领域，就像一块巨大的蛋糕，商家们各显神通，希望占据更多的市场份额，获得更大的利润。个别商家通过作弊手段希望获得更多利益，譬如：虚假交易就是一种重要的作弊方式，借此提升商家的等级，骗取用户的信任。不打击这些作弊的行为，就会极大地损害整个市场的信用体系，让诚信的商家蒙受损失，进而会有更多的商家尝试通过作弊来获取利益。作弊与反作弊的斗争一直在进行中，不断有新的方法出现，也不断有新的对策出台。

这里我们介绍一个例子，通过对交易行为的分析，预测商家作弊情况。注意：所使用的建模数据经过特殊处理，分析结论不能反映真实的交易情况。从机器学习方法的角度来看，这是典型的分类问题，而且分类目标为两个，使用的数据特征已经被很好地数字化，可以直接套用一些常用的分类模型进行训练、预测。

使用的数据表名为 business_fraud，有 1 个 ID 列，6 个属性列和 1 个标签列，各列的介绍如表 2-1 所示，各属性列如何变换到 0~1 区间，不是本节的重点，不展开讨论。

表 2-1　交易信息字段

列　名	含　义	备　注
Transaction_id	交易 ID	
a_score	账户价值	取值范围：[0，1]
b_score	购物类别	取值范围：[0，1]
r_score	消费频率	取值范围：[0，1]
p_score	注册时间	取值范围：[0，1]
ri_score	购物时间	取值范围：[0，1]
v_score	消费总额	
label	是否作弊	0-作弊操作，1-非作弊

数据如图 2-1 所示，很明显，字段 b_score、r_score 和 p_score 中 0 值的个数较多。

transaction_id	a_score	b_score	r_score	p_score	ri_score	v_score	label
0	0.177	0	0	0	0.373	0.019	0
1	0.002	0.919	0	0.311	0.517	0.329	0
2	0.362	0	0	0	0.195	0.02	0
3	0	0	0	0	0.37	0.176	0
4	0.808	0	0	0	0.236	0.667	0
5	0.002	0	0	0	0.462	0	0
6	0.718	0	0	0	0.152	0.422	0
7	0.818	0	0	0	0.174	0.565	0
8	0.711	0	0	0	0.118	0.279	0
9	0.01	0	0	0	0.279	0.199	0
10	0.709	0	0	0	0.259	0.34	0
11	0.135	0	0	0	0.253	0.096	0
12	0.848	0	0	0	0.093	0.705	0
13	0.012	0	0	0	0.279	0	0
14	0.085	0	0	0	0.512	0.002	0
15	0.189	0	0	0	0.419	0.007	0
16	0.847	0	0	0	0.066	0.456	0

图 2-1　交易信息数据表

2.1　数据探索

首先使用最常用的组件，“全表统计”和“直方图（多字段）”，关注最基本的统计信息。各组件的连接方式如图 2-2 所示。

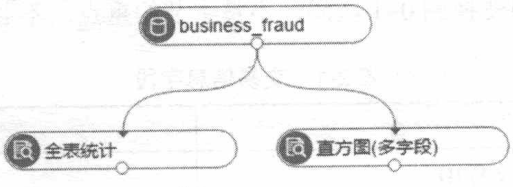


图 2-2　数据探索流程图

其中，“全表统计”组件可以使用默认参数，即对所有数据列进行统计；“直方图（多字段）”组件需要选择字段，如图 2-3 所示，在弹出的“选择字段”窗口选择所有特征属性列。