Universe Scientific Publishing
Computer System Networking and Telecommunications
计算机系统网络和电信，2019,1(2)
ISSN:2661-3719 (Print); 2661-3727 (Online)

On the Application of Data Mining Technology in Today's Era

Panpan NIU Zhengde BAO Yawen TANG
School of Computer and Software, Jincheng college, Sichuan University, Chengdu, Sichuan, 611731

Abstract

In the era of data, almost all the leading figures in the IT industry have their own huge data storage database. The rapid development of informatization is accompanied by the increasing amount of data to be analyzed, the larger the data to be analyzed and the more complex the operation of extracting useful information. This paper analyzes the application of data mining technology.

Key Words

Data Mining, Data Analysis, Data Processing, Data Application

DOI:10.18686/jsjxt.v1i2.682

论数据挖掘技术在当今时代的应用

钮盼盼 鲍正德 唐娅雯
四川大学锦城学院计算机与软件学院，四川成都，611731

摘 要

数据时代，几乎所有的 IT 行业领头大佬，都有着自己巨大数据存储的数据库。信息化的快速发展伴随要分析的数据量越来越大、分析的数据越大、提取有用信息的操作更加复杂，数据分析人员的工作量随着大大增加。本文以数据挖掘技术的应用为根本进行剖析。

关键词

数据挖掘；数据分析；数据处理；数据应用

1.引言

“大数据”在信息技术时代正迅速觉醒。与此同时，“大数据时代”紧随其后。“大数据时代”顾名思义，是一个全面呈现大数据的全新时代。大数据与我们生活息息相关的吃喝玩乐，甚至是与国家发展的医疗、科技、军事等重要领域都发挥着巨大的作用。结果，数据正在膨胀，甚至呈指数级爆炸。然而，最初产生的大量数据是一堆复杂的混乱的的数据，我们需要从中获得有用的数据则需要一定的技术，此时数据挖掘就在此扮演着重要的角色。挖掘有用的数据，并对数据进行分析从中得到一些决策性的信息，这将在商业、经济及其他领域中发挥着至关重要的作用。[1]正当的运用数据挖掘技术，得到真正需要的信息，能够更好的达到管理的目的。[2]

2.数据挖掘

大数据时代不仅是海量数据的爆发，其也在促进着各种处理数据的技术的发展。[3]与较早的数据处理技术做个比较，以前传统的数据处理技术大多是以统计分析为象征性的。以采用聚类、分类和关联分析为关键技术展开的数据挖掘主要在于发掘信息，挖掘出知识。从大量的复杂的无规律型的数据中获得数据能够产生的价值。无论是经济发展还是商业性，目前的数据挖掘技术都是对传统数据分析处理技术的一大优化。随着科技的发展,数据越来越庞大，数据处理技术的功能也会越来越强大。[4]

3.数据挖掘——分类分析

3.1 分类分析

分类分析简单的可以理解为就是将数据进行分类，但是分类得到的类别要有一定的依据，即分类的数据要具有大体相似的属性。一般而言，首先需要找到可以区分描述出数据类的模型，主要用建模的方法来推断出类

标记不知道的对象类。数据分类处理也可运用到数据预报中，其可以从历史数据中概括出对给予的数据来说明，从而得到对不了解数据的预测。分类通常输出的是离散型的数据，其体现大多以决策树的模式，按照你给的的数据值从树根循序扫描，沿着达到数据条件的分支进行上走，等到达树叶的时候急本就可以确定其类别了。除了决策树分析方法外，支持向量机（SVM）和贝叶斯网络也是及其关键的分析方法。

3.2 案例分析

现某银行有一张“客户信用卡申请信息.xlsx”表，用来研究客户能否申请到信用卡与那些因素有关，此表里面包含字段如下表 1 所示：

数据字段	描述	角色
编号	1000~	无
年龄	10~80	输入
性别	男　女	输入
户籍	北京，上海，…	输入
婚姻状况	丧偶，已婚，未婚，离异	输入
教育程度	初中及以下，大专，本科，硕士及以上，高中	输入
职业类型	个体户，其他企业，国有企业，外资企业，私营企业	输入
工作年限	0~50	输入
个人收入	10000~ 9.9E10	输入
居住类型	自购房，租房，其他	输入
车辆情况	有　无	输入
保险缴纳	有　无	输入
信贷情况	正在偿还 正常还款 没有贷款记录 现在没有贷款 还在拖欠 逾期还款	输入
信用等级	A~F	输入
是否申请成功	成功　失败	目标

表 1 客户信用卡申请信息

现以“是否申请成功”为目标字段，“编号”为无，其余为输入字段。以贝叶斯网络模型为分析的模型，经过“源”进行数据源的输入，再经过“类型”操作，选择“贝叶斯网络模型”可以得到如下图 1 结果：

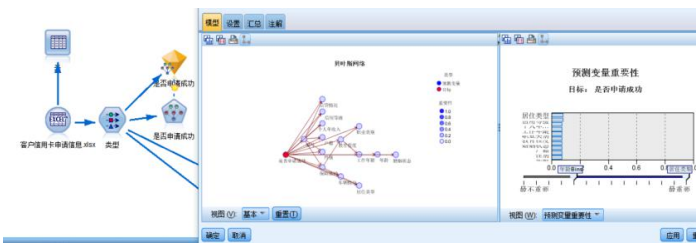

图 1 贝叶斯网络模型

对图 1 贝叶斯网络模型结果可以分析得到，影响客户是否能够申请到信用卡的最为关键因素为“信用等级”和“居住类型”。因此顾客在申请信用卡时，银行会着重考虑客户的信用等级和居住类型。

4.数据挖掘——聚类分析

4.1 聚类

聚类分析是在不清楚分类分析的前提下，按照信息具有的属性对信息进行聚在一起的一种数据处理形式。分类与聚类简单的理解定义可以是，依照数据的某些属性把数据分为有差异的类别即是分类，聚类则是根据数据的一些特征把数据聚为不同的类别（一分一聚，前者有监视性学习，后者无监视性学习）。实际运作中，聚类分析可以在大量的复杂的数据中，依据数据的某些特征，先对数据进行聚类分析，在将聚成的类别进行剖析。例如，在一份客户通话时长的数据表中，可以根据客户的不同通话性质（如会议通话、家庭通话等）、不同的通话时间段（工作日、周末、早、中、晚等），不同的通话时长（1~10h）等特征对客户进行聚类为商务用户、大型客户、普通用户等。然后可以按照聚类成的不尽相同的使用者，向其力荐不同的消费套餐。惯用的聚类方法有凝聚层次聚类，K 均值聚类算法等。

4.2 案例分析

现某银行有一张“是否存在欺诈.xlsx”表，以聚类分析研究哪些客户可能会存在欺诈的行为，此表 2 里面包含字段如下所示：

数据字段	描述	角色
编号	1000~	五
额度	10000~100000	输入
日均消费金额	30~81797	输入
日均次数	1~28	输入
单笔消费最大金额	30~500000	输入
个人收入——连续	10416~9.9E10	输入
是否存在欺诈	0　1	目标
单笔是否透支	超过　未超过	输入
日均消费是否超过收入	超过　未超过	输入
刷卡频率	不频繁，非常频繁，频繁	输入

表 2 是否存在欺诈

现在以“额度”，“日均消费金额”，“日均次数”，“单笔消费最大金额”和“个人收入”为关键字段进行操作处理。以 K 均值为此次的剖析模型，经过“源”进行数据源的输入，再经过“类型”操作，选择“K-Means 模型”将聚类数设置为“3”可以得到如下图 2 结果：

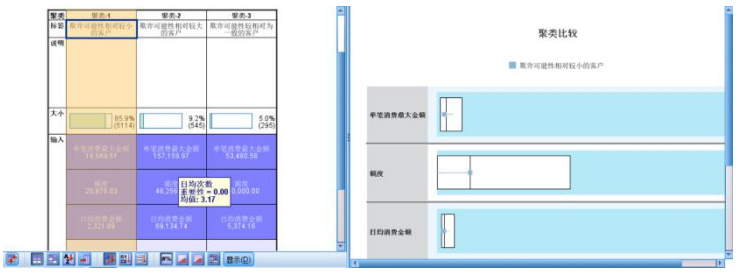

图 2 聚类模型

如图 2 聚类模型所示，依据“额度”，“日均消费金额”，“日均次数”，“单笔消费最大金额”和“个人收入”的指标可以将聚类 1~3 分为存在诈骗可能性较大，较小和一般的客户。

5.数据挖掘——关联分析

5.1 关联

关联分析就是在一堆看似杂乱无章的数据中，依据数据自身所具有的的某些特征属性，找到其可能潜在的联系，甚至依据这些联系会得到一些事物的规律性的东西。关联分析方法用以隐蔽的大型数据密集有趣的连接。往往由关联规则或频繁项集展示其联系。关联分析的算法有 Apriori 算法和 FP 增长算法。

Apriori 算法的中心思想：利用的层层迭代方法的基础上的候选人产生的频繁的项目组

其算法流程的方法概括可为“合并”与“剪枝”。合并：寻找最后一项相异的合并，剪枝：找到子集不是频繁项集的剪掉。

FP 增长算法是一种频繁模式增长算法，其思想是递归地增长频繁模式和数据库划定

例：下表 3 是一个涵盖了 7 个事务的事务集合。每个事务说明一位消费者在商店一次购物的货物集合。以最小支持度为 30%，最小置信度为 80%，按照 Apriori 算法找到所有的频繁项集和强关联性规则？

交易号码	商品列表
T001	牛肉，鸡肉，牛奶
T002	牛肉，奶酪
T003	奶酪，靴子
T004	牛肉，鸡肉，奶酪
T005	牛肉，鸡肉，衣服，奶酪，牛奶
T006	鸡肉，衣服，牛奶
T007	鸡肉，牛奶，衣服

表 3 商品表

解答：

一：找到全部的频繁项集（最小支持度为 30%）

1）1-项集：

1-候选项集 C_1 及其支持的：

sup{牛肉}=4/7，sup{鸡肉}=5/7，sup{牛奶}=4/7，sup{奶酪}=4/7，

sup{靴子}=1/7，sup{衣服}=3/7

1-频繁项集 F_1：

{牛肉}、{鸡肉}、{牛奶}、{奶酪}、{衣服}

2）2-项集：

2-候选项集 C_2 及其支持度：

sup{牛肉，鸡肉}=3/7，sup{牛肉，牛奶}=2/7，sup{牛肉，奶酪}=3/7，sup{牛肉，衣服}=1/7，sup{鸡肉，牛奶}=4/7，sup{鸡肉，奶酪}=2/7，sup{鸡肉，衣服}=3/7，sup{牛奶，奶酪}=1/7，sup{牛奶，衣服}=3/7，sup{奶酪，衣服}=1/7

2-频繁项集 F_2：

{牛肉，鸡肉}，{牛肉，奶酪}，{鸡肉，牛奶}，{鸡肉，衣服}，{牛奶，衣服}

3）3-项集：

3-候选项集 C_3 及其支持度：

sup{牛肉，鸡肉，奶酪}=2/7、sup{鸡肉，牛奶，衣服}=3/7

3-频繁项集 F_3：

{鸡肉，牛奶，衣服}

由上可得频繁项集有：

1-项集：{牛肉}、{鸡肉}、{牛奶}、{奶酪}、{衣服}

2-项集：{牛肉，鸡肉}，{牛肉，奶酪}，{鸡肉，牛奶}，{鸡肉，衣服}，{牛奶，衣服}

3-项集：{鸡肉，牛奶，衣服}

二：找到全部强关联规则（最小置信度为 80%）

1）由 2-构成关联规则及其置信度：

conf({牛肉}→{鸡肉})=3/4，conf({鸡肉}→{牛肉})=3/5，

conf({牛肉}→{奶酪})=3/4，conf({奶酪}→{牛奶})=3/4，

conf({鸡肉}→{牛奶})=4/5，conf({牛奶}→{鸡肉})=1，

conf({鸡肉}→{衣服})=3/5，conf({衣服}→{鸡肉})=1，

conf({牛奶}→{衣服})=3/4，conf({衣服}→{牛奶})=1

得到最小置信度为 80%强关联规则是：

{牛奶}→{鸡肉}，{鸡肉}→{牛奶}，{衣服}→{鸡肉}，{衣服}→{牛奶}

2）由 3-构成关联规则和其置信度：

conf({牛奶}→{鸡肉，衣服})=3/4，conf({牛奶，衣服}→{鸡肉})=1，conf({鸡肉}→{牛奶，衣服})=3/5，conf({鸡肉，牛奶}→{衣服})=3/4，conf({衣服}→{鸡

肉，牛奶})=1，conf({鸡肉，衣服}→{牛奶})=1，
得到最小置信度为 80%80%的强关联规则是：
{衣服}→{鸡肉，牛奶}，{鸡肉，衣服}→{牛奶}，{牛奶，衣服}→{鸡肉}

由上可得的强关联如下：
{牛奶}→{鸡肉}，{鸡肉}→{牛奶}，{衣服}→{鸡肉}，{衣服}→{牛奶}

{衣服}→{鸡肉，牛奶}，{鸡肉，衣服}→{牛奶}，{牛奶，衣服}→{鸡肉}

5.2 案例分析

现某银行有一张“欺诈人口属性分析.xlsx”表，图 3 里面包含如下图所示字段，现以此表来研究客户的“居住类型”与“车辆情况”之间的关联：

	居住类型	职业类别	工作年限	个人收入_连续	保险缴纳	车辆情况	是否存在欺诈
5935	租房	私营企业	19.000	216000000.0...	无	无	0.000
5936	租房	私营企业	0.000	216000000.0...	无	无	0.000
5937	其他	个体户	0.000	216000000.0...	无	无	0.000
5938	其他	私营企业	8.000	229000000.0...	有	无	0.000
5939	其他	国有企业	8.000	229000000.0...	有	无	0.000
5940	其他	外资企业	10.000	229000000.0...	有	无	0.000
5941	其他	私营企业	19.000	229000000.0...	有	无	0.000
5942	其他	私营企业	31.000	862400000.0...	无	无	0.000
5943	租房	私营企业	5.000	920000000.0...	无	无	0.000
5944	租房	国有企业	11.000	920000000.0...	无	无	0.000
5945	租房	私营企业	22.000	920000000.0...	无	无	0.000

图 3 欺诈人口属性

现以“居住类型”和“车辆情况”为关联字段，经过“源”对数据进行输入，在以“类型”进行操作后加入“设为标志”将居住类型和车辆情况选中创建为标志字段，随后在其后再次加入“类型”使“车辆情况_有/无”设置为目标，“居住类型_租房/自购房/其他”设置为输入，最后加入“Apriori 模型”可以得到图 4：

图 4 Apriori 模型

分析：如图 4 Apriori 模型在最低条件支持的为 0.0 和最小规则置信度（%）为 90.0 的条件下，可以得到，租房和其他居住类型的客户一般是没有车辆的，而自己拥有房子的客户是拥有车辆的，这也比较符合实际情况。

6.结束语

“数据爆炸”的是我们身处的时代。数据的产生促进着数据分析处理技术的发展。因此，许多的数据处理分析技术都开始慢慢发展起来。此时，对于上市公司的领导者来说，选取正当的数据剖析处理技术，从磅礴的数据中挖掘出深层次的信息，然后经过提炼得到真正对公司产生价值的数据至关重要。因而，数据挖掘技术在新时代新技术下出现了。[5]

参考文献

[1]周庆荣.数据挖掘技术在企业财务管理中的创新应用[J].科技创业月刊,2018,31(12):135-137.
[2]杨继武.大数据时代背景下数据挖掘技术的应用[J].电子技术与软件工程,2019(02):163.
[3]唐云凯,王芳,刘淑英.海量数据挖掘过程相关技术研究进展[J].电脑知识与技术,2018,14(36):1-2.
[4]曲萍.基于大数据时代的数据挖掘技术[J].中国新技术新产品,2019(04):20-21.
[5]梅拥军.软件工程中数据挖掘技术的应用[J].电子技术与软件工程,2019(01):141.

作者简介

第一作者：钮盼盼（1993-），女，汉，安徽省亳州市，本科，四川大学锦城学院，研究方向：大数据应用与商业智能。
第二作者（通讯作者）：鲍正德（1989-），男，汉，黑龙江哈尔滨，研究生，四川大学锦城学院，研究方向：电子商务。
第三作者：唐娅雯（1999-），女，汉，四川省资阳市，本科，四川大学锦城学院，研究方向：信息管理、J2EE