<<数据挖掘实用机器学习技术>>

图书基本信息

书名：<<数据挖掘实用机器学习技术>>

13位ISBN编号：9787111182054

10位ISBN编号：7111182057

出版时间：2006-3

出版时间：机械工业出版社

作者：Ian H.Witten,Eibe Frank

页数：362

译者：董琳,邱泉,于晓峰

<<数据挖掘实用机器学习技术>>

内容概要

本书介绍数据挖掘的基本理论与实践方法。
主要内容包括：各种模型(决策树，关联规则、线性模型、聚类、贝叶斯网以及神经网络)以及在实践中的运用，所存任缺陷的分析。
安全地清理数据集、建立以及评估模型的预测质量的方法，并且提供了一个公开的数据挖掘工作平台Weka。
Weka系统拥有进行数据挖掘仟务的图形用户界面，有助于理解模型，是一个实用并且深受欢迎的工具。
本书逻辑严密、内容翔实、极富实践性，适合作为高等学校本科生或研究生的教材，也可供相关技术人员参考。

作者简介

Ian H.Witten，新西兰怀卡托大学计算机科学系教授，ACM和新西兰皇（IFIP）颁发的Namur奖项。他的著作包括《Managing Gigabytes:Compressing and Indexing Documents and Images》、《How to Build a Digital Library》以及众多的期刊和学会文章。

<<数据挖掘实用机器学习技术>>

书籍目录

出版者的话专家指导委员会译者序中文版前言序前言第一部分 机器学习工具与技术 第1章 绪论 1.l 数据挖掘和机器学习 l.2 简单的例子：天气问题和其他 l.3 应用领域- 1.4 机器学习和统计学 1.5 用于搜索的概括 l.6 数据挖掘和道德 1.7 补允读物 第2章 输入概念、实例和属性 2.1 概念 2.2 样本 2.3 属性 2.4 输入准备 2.5 补充读物 第3章 输出：知识表达 3.1 决策表 3.2 决策树 3.3 分类规则 3.4 关联规则 3.5 包含例外的规则 3.6 包含关系的规则 3.7 数值预测树 3.8 基于实例的表达 3.9 聚类 3.10 补充读物 第4章 算法基本方法 4.1 推断基本规则 4.2 统计建模 4.3 分治法：创建决策树 4.4 覆盖算法：建立规则 4.5 挖掘关联规州 4.6 线性模型 4.7 基于实例的学习 4.8 聚类 4.9 补充读物 第5章 可信度：评估机器学习结果 5.1 训练和测试 5.2 预测性能 5.3 交叉验证 5.4 其他估计法 5.5 可信度：评估机器学习结果 5.6 预测概率 5.7 计算成本 5.8 评估数值预测 5.9 最短描述长度原理 5.10 聚类方法中应用MDL原理 5.1l 补充读物 第6章 实现：真正的机器学习方案 …… 第7章 转换：处理输入和输出 第8章 继续扩展和应用 第9章 Weka简介 第10章 Explorer界面 第1l章 Knowledge Flow界面 第12章 Experimenter界面 第13章 命令行界面 第14章 嵌入式机器学习 第15章 编写新学习方案参考文献索引

章节摘录

第7章 转换：处理输入和输出　　在前一章中我们考察了大量的机器学习方法：决策树、决策规则、线性模型、基于实例的方案、数值预测技术、聚类算法以及贝叶斯网络。
所有这些方法都是合理、成熟的技术，可用于解决实际的数据挖掘问题。
　　但是成功的数据挖掘远不只是牵涉到选择某种学习算法并应用于数据。
许多学习算法要用到各种不同的参数，需要选择合适的参数值。
在多数情况下，选择适当的参数可以使所获结果得到显著改善，而合适的选择则是要视手头的具体数据而定的。
例如，决策树可以选择修剪或不修剪，选择前者又需要选择修剪参数。
在基于实例的k最近邻学习方法中，则需要选择k值。
更为常见的，则是需要从现有的方案中选择学习方法本身。
在所有情况下，合适的选择是由数据而决定的。
　　在数据上试用几种不同的方法，并使用几种不同的参数值，然后观测哪种情况结果最好，是个诱人的方法。
不过要当心！
最佳选择并不一定是在训练数据上获得最好结果的那个。
我们曾反复提醒要注意过度拟合问题，过度拟合是指一个学习模型与用于建模的某个具体训练数据集太过匹配。
假设在训练数据上所表现的正确性能代表模型将来应用于实践中的新数据上的性能水准，这个想法是不正确的。
　　所幸的是在第5章中已经讨论了对于这个问题的解决方法。
有两种较好的方法可用来估计一个学习方法的预期真实性能表现：在数据源充足的情况下，使用一个与训练数据集分离的大数据集；在数据较少的情况下则使用交叉验证法（第5.3节）。
在后一种情况下，在实践中的典型应用方法是单次的10折交叉验证，当然要得到更为可靠的估计需要将整个过程重复10次。
一旦为学习方法选定了合适的参数，就可以使用整个训练集（即所有训练实例）来生成将要应用于新数据的最终学习模型。
　　注意在调整过程中使用所选的参数值得到的性能表现并不是对最终模型性能的一个可靠估计，因为最终模型对于调整中使用的数据有过度拟合的倾向。
要确定它的性能究竟如何，需要另外一个大的数据集，这个数据集须与学习过程和调整过程中所使用的数据隔离开来。
在进行交叉验证时也是如此，参数调整过程需要一个“内部”交叉验证，误差估计还需要一个“外部”交叉验证。
采用10折交叉验证法将使学习方法运行100次。
总而言之，当评估一个学习方案的性能时，所进行的任何参数调整过程都应被看作是训练过程的一部分。
　　当把机器学习技术应用于实际的数据挖掘问题时，还有其他一些重要程序可以大大提高成功率，这正是本章的主题。
它们形成了一种（操纵）数据的技术，将输入数据设计成一种能适合所选学习方案的形式，将输出模型设计得更为有效。
你可以把它们看成是能应用于实际的数据挖掘问题以提高成功几率的一些诀窍。
有时奏效，有时无效。
根据目前的技术发展水平来看，很难预言它们是否有用。
在这种以尝试和误差率作为最为可靠的指导的领域中，特别重要的恐怕就是灵活运用并且理解这些诀窍了。
　　……

编辑推荐

正如所有受到商业注目的新兴技术一样，数据挖掘的运用也是极其多样化的。
言过其实的报导声称可以建立算法，在数据的海洋里发现秘密。
但事实上机器学习中没有魔术，没有隐藏的力量，没有炼金术。
有的只是一些可以将有用的信息从原始数据中提炼出来的清晰明了的实用技术。
《数据挖掘实用机器学习技术》(原书第2版)叙述了这些技术并展示了它们是如何工作的。
　　《数据挖掘实用机器学习技术》(原书第2版)对1999年的初版做了重大的改动。
虽说核心概念没有变化，但《数据挖掘实用机器学习技术》(原书第2版)做了更新，反映出过去五年的变化。
《数据挖掘实用机器学习技术》(原书第2版)提供了机器学习理论概念的完整基础，此外还对实际工作中应用的相关工具和技术提了一些建议。
　　《数据挖掘实用机器学习技术》(原书第2版)逻辑严密、内容翔实、极富实践性，适合作为高等学校本科生或研究生的教材，也可供相关技术人员参考。

版权说明