Data Mining: Introduction

Lecture Notes for Chapter 1

Introduction to Data Mining, 2nd Edition

by

Tan, Steinbach, Karpatne, Kumar

大规模数据无处不在!
有大量的数据
商业和商业的增长
科学数据库由于
数据生成的进展
和收集技术
新咒语
尽可能收集数据
无论何时何地
有可能。
预期
收集的数据会有价值
不管是为了什么目的
收集的或者不是为了某个目的
预想的。
计算模拟
社交网络:推特
传感器网络
交通模式
网络安全
01/17/2018 数据挖掘导论，第 2 版 2
电子商务

为什么是数据挖掘？商业观点

●正在收集大量数据
和仓库
网络数据
uYahoo 拥有千兆字节的网络数据
uFacebook 拥有数十亿活跃用户
部门/部门的采购
杂货店，电子商务
亚马逊每天处理数百万的访问量
银行/信用卡交易
●计算机变得更便宜、更强大
●竞争压力很大
为边缘提供更好的定制服务(例如
客户关系管理)

为什么是数据挖掘？科学观点
●数据收集和存储在
巨大的速度
卫星上的遥感器
美国宇航局 EOSDIS 档案馆结束
每年 1pb 的地球科学数据
望远镜扫描天空
天空调查数据
高分辨率生物数据
科学模拟
在几个小时内生成了 u 兆字节的数据
●数据挖掘帮助科学家
在大规模数据集的自动分析中
在假设形成中
01/17/2018 数据挖掘导论，第 2 版 4
脑天空调查数据的功能磁共振成像数据
基因表达数据
地球表面温度

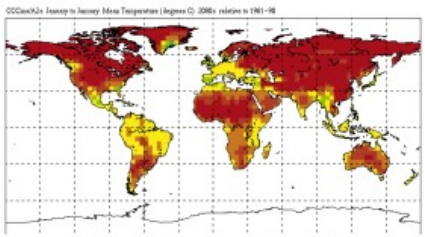
Great opportunities to improve productivity in all walks of life

McKinsey Global Institute

Big data: The next frontier for innovation, competition, and productivity

Great Opportunities to Solve Society's Major Problems

Improving health care and reducing costs



Predicting the impact of climate change

Finding alternative/ green energy sources

Reducing hunger and poverty by increasing agriculture production

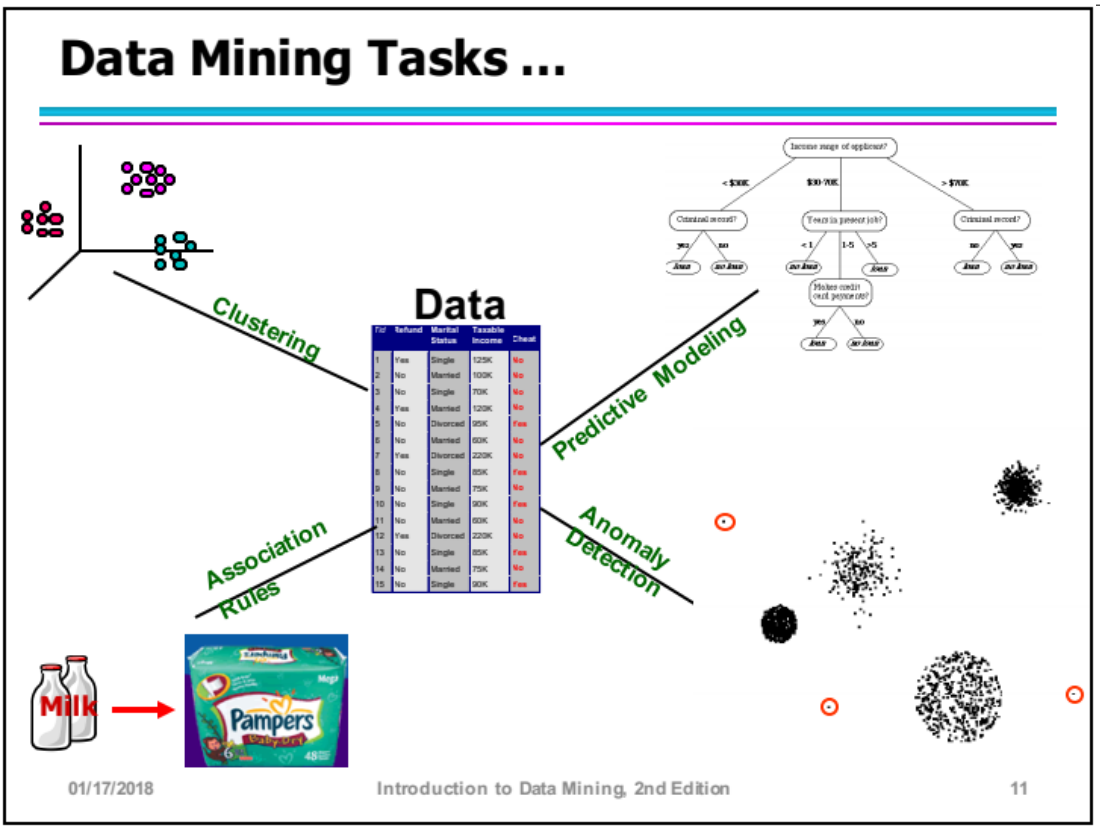
什么是数据挖掘?
●许多定义
隐含的、先前未知的非平凡提取
以及来自数据的潜在有用信息
探索和分析，自动或半自动
意味着，为了发现大量的数据
有意义的模式

什么是(不是)数据挖掘?
●什么是数据挖掘?
某些名字更多
在某些美国流行
地点(奥布赖恩、奥鲁克、
奥赖利…在波士顿地区)
相似的组合在一起
退回的文件
搜索引擎根据
他们的背景(例如亚马逊
雨林，Amazon.com)
●什么不是数据
采矿?
查找电话
电话号码
目录
查询网页
搜索引擎
关于...的信息
“亚马逊”

●从机器学习/人工智能、模式识别、
统计和数据库系统
●传统技术可能不适合，因为数据
大规模的
高维
异种的
复杂的
分布式的
●新兴数据科学和数据领域的一个关键组成部分-
驱动发现
数据挖掘的起源

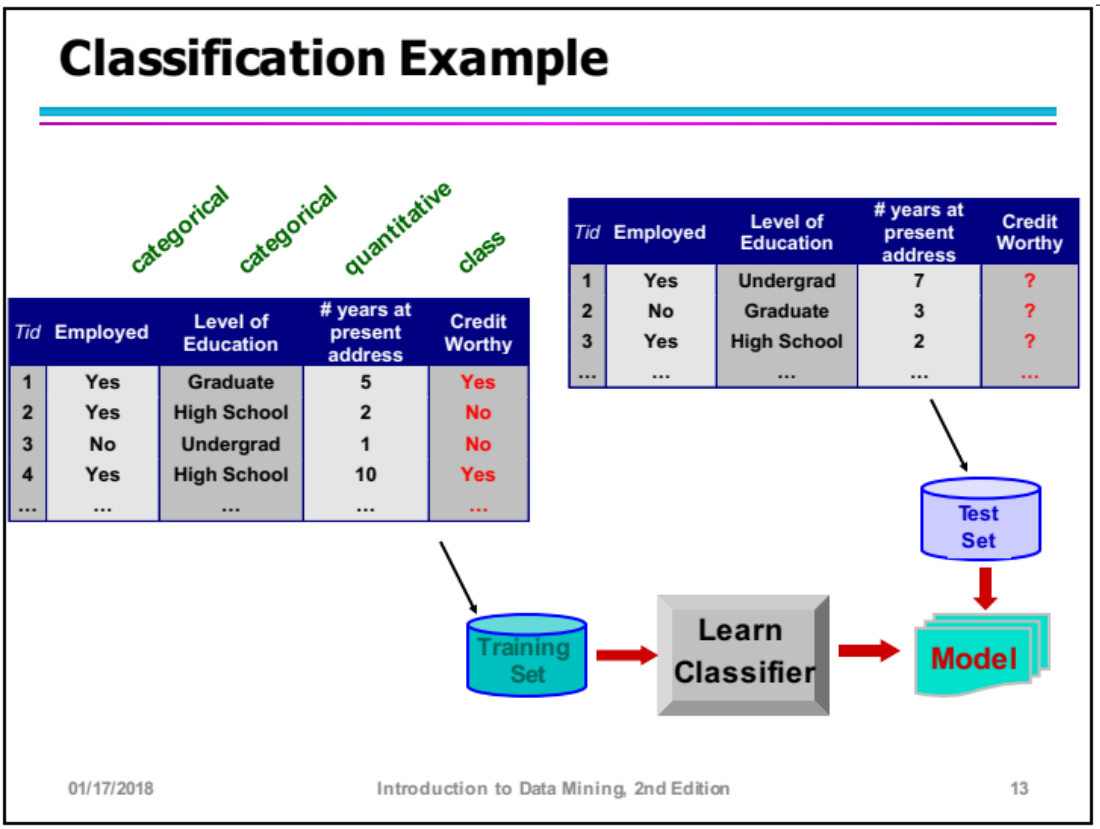
数据挖掘任务

●预测方法
使用一些变量来预测未知或
其他变量的未来值。
●描述方法
寻找人类可解释的模式
描述数据。
摘自[·法耶兹等著《知识发现和数据挖掘的进展》，1996 年
01/17/2018 数据挖掘导论，第 2 版 10



●为类属性找到一个模型，作为
其他属性的值
就业教育水平#年目前解决信用价值
1 是毕业生 5 是
2 是高中 2 不是
3 号大学生 1 号
4 是的高中 10 是的
… … … 10
信用预测模型
值得
就业阶层
没有教育
数量
年
不，是的

毕业{高中，本科}
是不是
> 7 岁< 7 岁
是
数量
年
不
> 3 年< 3 年
预测建模:分类
01/17/2018 数据挖掘导论，第 2 版 12



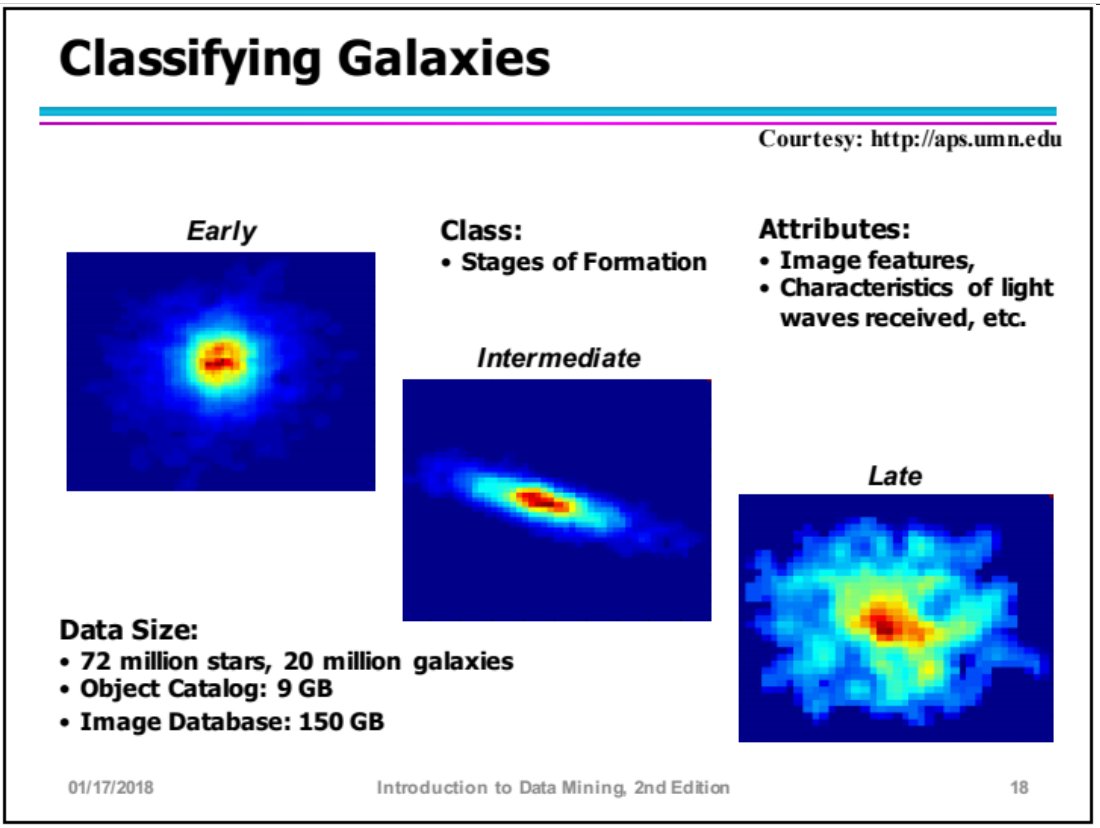
●信用卡交易分类
合法或欺诈
●土地覆盖分类(水体、城市地区、森林等)。)使用卫星数据
●将新闻故事归类为金融，
天气、娱乐、运动等
●识别网络空间中的入侵者
●预测肿瘤细胞是良性还是恶性
●分类蛋白质的二级结构
如 α 螺旋、β 片或随机线圈
分类任务示例
01/17/2018 数据挖掘导论，第 2 版 14

分类:应用 1

●欺诈检测
目标:预测信用卡欺诈案件
交易。
方法:
u 使用信用卡交易和信息
作为属性的帐户持有者。
顾客什么时候买，买什么，多长时间按时付款，等等
u 将过去的交易标记为欺诈或公平
交易。这形成了类属性。
u 学习交易类的模型。
u 使用此模型通过观察信用来发现欺诈
账户上的卡交易。
01/17/2018 数据挖掘导论，第 2 版 15

分类:应用 2
●电话客户的流失预测
目标:预测客户是否有可能
输给竞争对手。
方法:
u 使用每个交易的详细记录
过去和现在的客户，寻找属性。
顾客多久打一次电话，他打电话到哪里，什么时间打电话最多，他的财务状况，婚姻状况等。
给顾客贴上忠诚或不忠诚的标签。
找到忠诚的模式。
来自[·贝里和林诺夫]数据挖掘技术，1997
01/17/2018 数据挖掘导论，第 2 版 16

分类:应用 3
●天空调查编目
目标:预测天空物体的类别(恒星或星系)，
尤其是视觉模糊的，基于望远镜
勘测图像(来自帕洛马天文台)。
3000 幅图像，每幅图像 23，040 x 23，040 像素。
方法:
u 分割图像。
u 测量图像属性(特征)每 40 个
对象。
基于这些特征对类建模。
u 成功故事:能发现 16 个新的高红移
类星体，一些最遥远的困难物体
去寻找！摘自[·法耶兹等著《知识发现和数据挖掘的进展》，1996 年
01/17/2018 数据挖掘导论，第 2 版 17

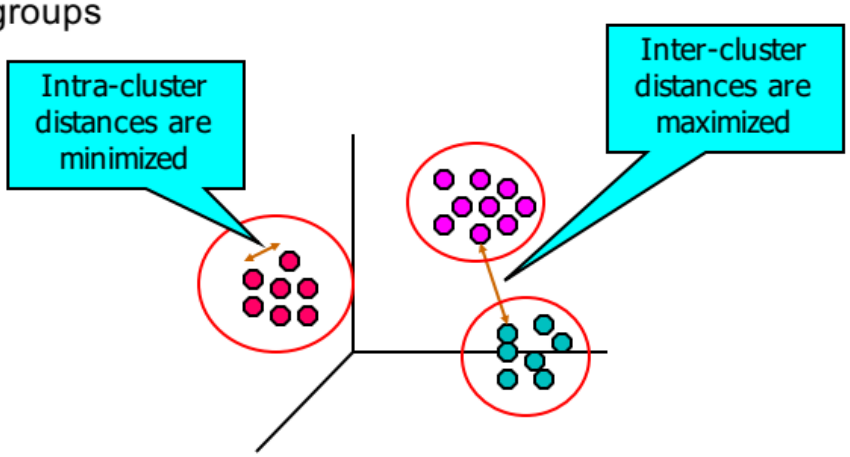


回归
●预测给定连续值变量的值
基于其他变量的值，假设
依赖的线性或非线性模型。
●在统计学、神经网络领域广泛研究。
●示例:
基于以下因素预测新产品的销售额
宣传支出。

预测风速作为
温度、湿度、气压等。
股票市场指数的时间序列预测。
01/17/2018 数据挖掘导论，第 2 版 19

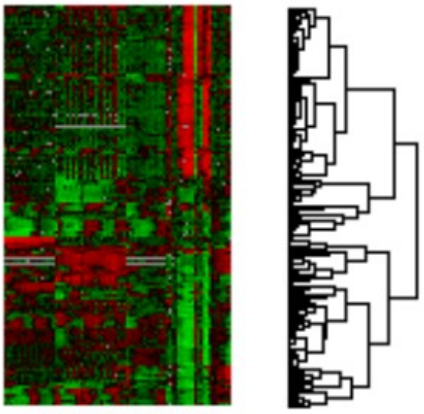
Clustering

- Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups

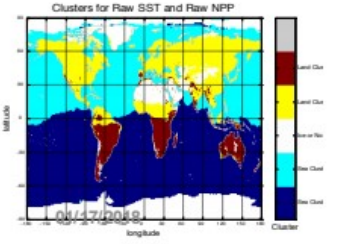


Applications of Cluster Analysis

- **Understanding**
  - Custom profiling for targeted marketing
  - Group related documents for browsing
  - Group genes and proteins that have similar functionality
  - Group stocks with similar price fluctuations
- **Summarization**
  - Reduce the size of large data sets



Courtesy: Michael Eisen



Use of K-means to partition Sea Surface Temperature (SST) and Net Primary Production (NPP) into clusters that reflect the Northern and Southern Hemispheres.

集群:应用 1
●市场细分:
目标:将市场细分为不同的子集
任何子集都可能存在的客户
被选为市场目标
独特的营销组合。
方法:
基于以下因素收集客户的不同属性
他们的地理和生活方式的相关信息。
u 寻找相似客户群。
通过观察购买来衡量集群质量
同一集群中的客户模式与
来自不同的集群。
01/17/2018 数据挖掘导论，第 2 版 22

集群:应用 2
●文档聚类:
目标:查找与相似的文档组
基于出现在
他们。
方法:识别中经常出现的术语
每份文件。基于形成相似性度量
不同术语的频率。用它来聚类。
数据挖掘导论，第 2 版 23
Enron 电子邮件数据集
2018 年 17 月 1 日

关联规则发现:定义
●给定一组记录，每个记录包含
给定集合中的一些项目
产生依赖规则，该规则将预测
基于其他项目的出现的项目的出现
物品。
TID 项目
1 面包，可乐，牛奶 2 啤酒，面包
3 啤酒，可乐，尿布，牛奶 4 啤酒，面包，尿布，牛奶 5 可乐，尿布，牛奶
发现的规则:
{牛奶} > {可乐}
{尿布，牛奶} > {啤酒}
01/17/2018 数据挖掘导论，第 2 版 24

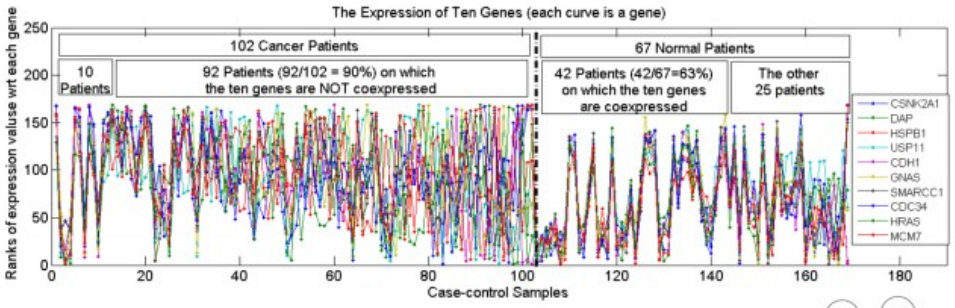
关联分析:应用
●市场篮分析
规则用于促销、货架
管理和库存管理

●电信报警诊断
规则用于查找
在同一时期经常一起出现
●医学信息学
规则被用来寻找病人的组合
症状和测试结果与某些
疾病
01/17/2018 数据挖掘导论，第 2 版 25

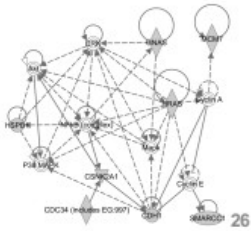
Association Analysis: Applications

- An Example Subspace Differential Coexpression Pattern from lung cancer dataset

Three lung cancer datasets [Bhattacharjeeet 2001], [Stearman et al. 2005], [Su et al. 2007]



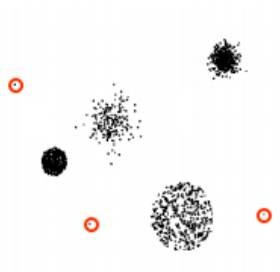
Enriched with the TNF/NFB signaling pathway which is well-known to be related to lung cancer
P-value: $1.4*10^{-5}$ (6/10 overlap with the pathway)

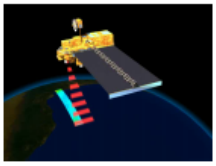


[Fang et al PSB 2010]

Deviation/Anomaly/Change Detection

- Detect significant deviations from normal behavior
- Applications:
  - Credit Card Fraud Detection
  - Network Intrusion Detection
  - Identify anomalous behavior from sensor networks for monitoring and surveillance.
  - Detecting changes in the global forest cover.





Motivating Challenges

- Scalability
- High Dimensionality
- Heterogeneous and Complex Data
- Data Ownership and Distribution
- Non-traditional Analysis