数据挖掘

- 任务
  - 分类预测型任务
    - 分类和预测
    - 预测回归
    - 异常发现
  - 描述型任务
    - 聚类
    - 摘要
    - 依赖关系
- 数据预处理
  - 概述
    - 重要性
      - 高质量的数据才能产生高质量的挖掘结果
      - 数据预处理是数据挖掘过程中占工作量最大的一个步骤
    - 要求
      - 精确度
      - 完整度
      - 一致性
      - 可信度
      - 附加价值
      - 可解释性
    - 主要任务
      - 数据清理
      - 数据集成
      - 数据变换
      - 数据归约
  - 数据对象及其统计描述
    - 数据对象
      - 数据集
      - 数据对象
    - 属性
      - 一般数据库的行对应数据对象，列对应属性
      - 属性类型
    - 数据的统计描述
      - 度量数据的中心趋势
        - 平均值
        - 中位数
        - 中列数
        - 众数
        - 对称数据
        - 倾斜数据
          - 正倾斜
          - 负倾斜
      - 度量数据的离散趋势
        - 极差
        - 四分位数 — IQR = Q3-Q1
        - 离群点 — 挑出至少高于第三个四分位数或低于第一个四分位数1.5*IQR处的值
        - 盒图
        - 方差和标准差
      - 相关的图形显示
    - 数据相似性和相异性
      - 数据矩阵
      - 相异性矩阵
      - 标称属性的邻近性度量
        - 简单匹配方法
        - 创建二元变量，使用jaccard系数，本质上和简单匹配法很相似
      - 二元属性的邻近性度量（两个计算有一点小小的差别）
        - 对称属性 — 属性两个状态具有同等价值，相同权重
        - 非对称属性 — 属性两状态重要性不同
      - 数值属性的临近性度量 — 就是算两个对象间的距离，那几个距离公式
      - 序数属性的临近性度量 — 序数属性的值之间有特定意义，度量还蛮复杂
      - 向量对象的临近性度量 — 余弦相似性，自然语言处理中应用超级广泛
    - 数据清理
      - 填写空缺的值
      - 识别离群点和平滑噪声数据 — 有个分箱的概念
      - 纠正不一致的数据
      - 解决数据集成造成的冗余
    - 数据集成和变换
      - 数据集成 — 合并多个数据源中的数据，存放在一个一致的数据存储中，这当中要做到一致性
      - 数据变换
        - 将数据转换成适合挖掘的统一形式
        - 最小-最大规范化
        - z-score规范化 — 最大最小值不知道时适用
    - 数据归约
      - 常用数据归约策略
        - 维归约：移除不重要的属性
        - 主成分分析PCA等等
      - 数据压缩
        - 有损压缩
        - 无损压缩
      - 数据归约
        - 使用模型来表示数据，通过选择替代的或者较小的数据表示形式来减少数据量。
        - 有参方法 — 用参数模型估计，最后只用存储参数
          - 线性回归
          - 多元回归
        - 无参方法
          - 直方图
          - 聚类
          - 采样 — 用较少的随机样本来表示原始数据集
- 关联分析
  - 概述
    - 用于发现隐藏在大型数据集中令人感兴趣的频繁出现的模式、关联和相关性
    - 关联规则可以为市场预测、决策和策划等方面提供依据。
  - 基本概念和解决方法
    - 项与项集
      - 项集的支持度 — 某个项集出现在事务集中的概率，含有的数量/总的数量
      - 项集的最小支持度与频繁集
        - 用于发现关联规则的项集必须满足的最小支持度的阈值，称为项集的最小支持度(Minimum Support)，记为sup_min.满足最小支持度的项集才能产生关联规则
        - 支持度≥sup_min的项集称为频繁项集，简称频繁集，反之称为非频繁集
    - 事物
      - 每个事务T是项集I上的一个子集
      - 事务的全体被称为事务数据集，记作事务集
      - ■ 关联规则(Association Rule)是一个蕴涵表达式：
        R：$A \Rightarrow B\ [s, c]$
        - 其中$A \subset I$，$B \subset I$，并且$A \cap B = \varnothing$
        - 例如：规则R_1：{bread} ⇒ {milk}；规则R_2：{cream} ⇒ {bread, milk}；它们都可能是用户感兴趣的关联规则。
      - ■ 关联规则的度量：支持度s和置信度c。
    - 关联规则
      - 关联规则的支持度
        - 指事务集D中同时包含A和B的事务数和所有事务数之比
        - 反映了A和B所含的项在事务集中同时出现的概率
      - 关联规则的置信度
        - D中包含A和B的事务数与包含A的事务数的商
        - 反映了如果事务中包含A，则事务同时出现B的概率
      - 关联规则的最小支持度 — 关联规则需要满足的最低支持度，sup_min
      - 关联规则的最小可信度 — 关联规则需要满足的最低可信度，conf_min
      - 强关联规则
        - 同时满足大于关联规则的最小支持度和最小可信度称为强关联规则
        - 一般强关联规则才可以用于指导决策
    - 关联规则挖掘
      - 发掘频繁项集：事务支持度大于预先给定的最小阈值的项的集合。简单讲就是找到出现次数比较多的项集。
      - 产生关联规则 — 使用频繁项集来产生数据库中置信度c大于预先给定的最小阈值的关联规则
      - 主要挑战 — 会产生大量满足最小支持度的项集，这样的话频繁项集经过排列组合会有很多的项集，很难计算和存储
    - 闭频繁项集
      - 频繁闭项集 — 这个可以理解为在频繁项集的基础上按照出现的频率依次找，找不被别人包括的项集作为频繁闭项集
      - 极大频繁闭项集 — 在频繁项集里面找到最大的也就是包含最多的频繁项
      - $S = \{A, B, C, D\}$，$sup_{min} = 2$。
        - 1)频繁项集$\mathcal{F}$
          $\mathcal{F}$={<A>:3, <B>:3, <C>:3,<A,B>:3,<A,C>:2,<B,C>:2,<A,B,C>:2}
        - 2)频繁闭项集$\mathcal{C}$
          $\mathcal{C}$={<C>:3,<A,B>:3},<A,B,C>:2}
        - 3)极大频繁项集$\mathcal{M}$
          $\mathcal{M}$={<A,B,C>:2}}
        - 观察：$\mathcal{M} \subseteq \mathcal{C} \subseteq \mathcal{F}$，然而$\mathcal{C}$可以显著减少模式数量，保持了频繁项集的完整信息，由频繁闭项集可获得所有的频繁项集和它们的支持度，因此挖掘闭项集是一个好的选择。

TID	购买的item
1	C,D
2	A,B
3	A,B,C
4	A,B,C

  - 经典算法
    - 关联规则挖掘的两个步骤
      - 找出所有频繁集
      - 由频繁集产生强关联规则
    - Apriori算法
      - 通过迭代来穷举出数据集中的所有频繁集
      - 首先产生1-频繁集，然后通过连接和修剪产生2-频繁集，依次进行下去，直到最后无法产生新的频繁集。
      - 连接
      - 修剪
      - 该算法的缺点以及如何提高该算法的效率
    - FP-growth算法 — 该算法用到频繁模式树FP-tree，后期进行专门学习！
- 文本和Web挖掘
  - 文本信息检索
    - 信息检索模型
      - 布尔模型
      - 向量空间模型
      - 概率模型
    - 性能评价
      - 查全率
      - 查准率
    - 基于相似性的检索
  - 文本分类 — 利用计算机将一篇文档自动分配到一个或多个预定义的类别中
  - 文本挖掘
    - 自动摘要 — 就是对文档处理得到摘要啊
      - 分类
        - 主题摘要
        - 摘录型摘要
        - 评论型摘要
      - 步骤
        - 文档预处理
        - 过滤
        - 分词
        - 统计分析
        - 提取摘要
        - 输出摘要
        - 评估摘要
  - web挖掘
    - 分类
      - web内容挖掘
      - web使用挖掘
        - 从服务器的log文件或其他数据中分析用户的访问模式
        - 采用手段
          - 关联规则挖掘
          - 聚类分析
        - 阶段
          - 1.数据预处理阶段
          - 2.数据挖掘阶段
          - 3.结果分析阶段
      - web结构挖掘
        - 潜在的链接结构模式
        - page-rank算法
- 分类和预测
  - 决策树
    - 决策树的生成算法
      - ID3：信息增益 — 只能离散值
      - C4.5：信息增益比 — 离散+连续；连续值的离散化策略和方法
    - 决策树的剪枝策略
      - 预剪枝
      - 后剪枝
  - 神经网络
    - 概述
      - 三起三落
      - 四种分类
    - 神经元
      - 一个神经元又多个输入一个输出、一个内部反馈和阈值组成的非线性单元
      - 激活函数
    - 多层感知机网络
      - 网络结构
      - BP算法
        - 正向阶段
        - 反向阶段
  - 贝叶斯网络
    - 概率基础
      - 先验概率
      - 后验概率
      - 条件概率
      - 全概率公式 — 由原因推结果
      - 贝叶斯公式(后验概率公式) — 由结果推原因
    - 贝叶斯网络概述
      - 描述随机变量之间依赖关系。通过有向图表示随机变量间的因果关系，通过条件概率将因果关系量化
      - 构成
        - 网络结构
        - 条件概率表
    - 三个重要问题
      - 贝叶斯网络预测
        - 已知一定的原因，利用贝叶斯网络进行计算，由原因到结果的概率
        - 自顶向下的过程
        - 预测算法
      - 贝叶斯网络诊断
        - 发生了结果，找到该结果发生的原因以及发生的概率
        - 自底向上过程，全概率公式和贝叶斯公式
        - 诊断算法
      - 贝叶斯网络学习
        - 利用现有数据对先验知识进行修正的过程
        - 贝叶斯网络的建立
          - 1.抽象问题转化为网络中的节点。
          - 2.两个或多个结点之间建立连线
        - 贝叶斯网络的训练
          - 结点的概率
          - 节点间的条件概率
          - 多个节点间的联合条件概率
  - SVM
    - 概述
      - 以训练误差作为最优的约束条件，以置信范围最小作为优化目标，是一种基于结构风险最小化准则的学习方法，推广能力明显优于传统学习方法
    - 基本原理
      - 线性可分 — 分类间隔最大化
      - 线性不可分
        - 1.引入松弛变量构建软间隔SVM。 离群点。
        - 2.引入核函数，转到高维空间寻找分类超平面
    - 多分类问题 — 两种解决办法
    - SVM工具
      - libsvm
      - svm^light
  - 无监督：聚类分析
    - 聚类分析概述
      - 聚类分析作用
        - 获取数据集中数据的分布情况
        - 作为其它数据挖掘算法的预处理步骤
      - 方法
        - 划分法 — 以距离作为数据集中不同数据间的相似性度量，将数据集划分成多个簇
        - 层次法 — 对给定数据集进行层次划分，形成一个树形的聚类结果
    - 相似性计算方法
      - 连续型属性的相似性计算方法
        - 欧氏距离
        - 曼哈顿距离
        - 闵可夫斯基距离
      - 二值离散型属性的相似性计算方法
        - 可能性矩阵
        - 简单匹配系数(SMC)：对称二值离散型属性
        - Jaccard系数(JC)：不对称二值离散型属性
      - 多值离散型属性的相似性计算方法
        - 简单匹配法(SMC)
        - 将多值离散型属性转换成多个二值离散型属性，再使用jaccard系数计算样本之间的距离
      - 混合类型属性的相似性计算方法
        - 将混合类型预处理转化为只包含连续型属性和二值离散型属性的，给每个属性顶搞一个权重就可以解决问题
    - 聚类算法
      - 划分方法
        - k-means算法
        - k-medoids算法：不算均值，用数据点作为中心值
      - 层次方法
        - 合并聚类：AGNES算法
        - 分裂聚类：DIANA算法
        - CURE算法
      - 基于密度方法
    - 孤立点分析
      - 问题
        - 给定数据集中什么样的数据可以被认为是不一致的孤立点
        - 有效的方法挖掘孤立点
      - 常用方法
        - 基于统计
          - 一元正态分布，落在2.5%以外的样本可视为孤立点
          - 通常只能针对单一属性进行分析，且大多情况数据集概率分布未知
        - 基于距离
  - 结构风险最小化原理 — 经验风险和置信风险的和最小