信息科学技术学术著作丛书

数据约简——样例约简与属性约简

翟俊海 著

科学出版社

信息科学技术学术著作丛书

数据约简——样例约简与属性约简

翟俊海 著

科学出版社

北京

内 容 简 介

数据约简包括样例约简和属性约简，是从不同角度对数据进行约简. 本书在分类的框架下介绍数据约简的方法，重点介绍了确定性与不确定性环境下的样例约简方法和属性约简方法. 样例约简方法包括交叉选择样例算法、压缩模糊 K 近邻规则方法、概率神经网络样例选择算法. 属性约简方法包括最小相关性最大依赖度属性约简方法、模糊属性约简方法及属性约简方法在模型选择中的应用. 另外，本书还介绍了样例选择准则和特征子集评价准则. 本书以监督学习的基本理论为基础，全面系统地讨论了数据约简中的主要问题.

本书可作为应用数学、计算机科学与技术、自动化等专业高年级本科生和研究生的教材，也可供从事相关研究工作的科研人员参考.

图书在版编目(CIP)数据

数据约简：样例约简与属性约简/翟俊海著. —北京：科学出版社，2015
(信息科学技术学术著作丛书)

ISBN 978-7-03-044096-9

Ⅰ. 数… Ⅱ. 翟… Ⅲ. 数据采集 Ⅳ. TP274

中国版本图书馆 CIP 数据核字(2015)第 076934 号

责任编辑：魏英杰／责任校对：郭瑞芝
责任印制：张 倩／封面设计：陈 敬

科学出版社 出版
北京东黄城根北街 16 号
邮政编码：100717
http://www.sciencep.com

文林印务有限公司 印刷
科学出版社发行 各地新华书店经销
*
2015 年 5 月第 一 版 开本：B5(720×1000)
2015 年 5 月第一次印刷 印张：14 3/4
字数：298 000

定价：80.00 元
(如有印装质量问题，我社负责调换)

《信息科学技术学术著作丛书》序

21 世纪是信息科学技术发生深刻变革的时代, 一场以网络科学、高性能计算和仿真、智能科学、计算思维为特征的信息科学革命正在兴起. 信息科学技术正在逐步融入各个应用领域并与生物、纳米、认知等交织在一起, 悄然改变着我们的生活方式. 信息科学技术已经成为人类社会进步过程中发展最快、交叉渗透性最强、应用面最广的关键技术.

如何进一步推动我国信息科学技术的研究与发展; 如何将信息技术发展的新理论、新方法与研究成果转化为社会发展的新动力; 如何抓住信息技术深刻发展变革的机遇, 提升我国自主创新和可持续发展的能力? 这些问题的解答都离不开我国科技工作者和工程技术人员的求索和艰辛付出. 为这些科技工作者和工程技术人员提供一个良好的出版环境和平台, 将这些科技成就迅速转化为智力成果, 将对我国信息科学技术的发展起到重要的推动作用.

《信息科学技术学术著作丛书》是科学出版社在广泛征求专家意见的基础上, 经过长期考察、反复论证之后组织出版的. 这套丛书旨在传播网络科学和未来网络技术, 微电子、光电子和量子信息技术、超级计算机、软件和信息存储技术, 数据知识化和基于知识处理的未来信息服务业, 低成本信息化和用信息技术提升传统产业, 智能与认知科学、生物信息学、社会信息学等前沿交叉科学, 信息科学基础理论, 信息安全等几个未来信息科学技术重点发展领域的优秀科研成果. 丛书力争起点高、内容新、导向性强, 具有一定的原创性;体现出科学出版社“高层次、高质量、高水平”的特色和“严肃、严密、严格”的优良作风.

希望这套丛书的出版, 能为我国信息科学技术的发展、创新和突破带来一些启迪和帮助. 同时, 欢迎广大读者提出好的建议, 以促进和完善丛书的出版工作.

中国工程院院士

原中国科学院计算技术研究所所长

前　言

随着计算机技术、数据存储技术、无线通信技术等的快速发展, 数据的获取和存储都变得相对容易而且廉价. 对用户来说, 特别是企业用户, 存储的数据具有巨大的潜在应用价值, 数据所起的作用越来越显著. 研究如何从数据中高效地学习 (或挖掘) 有价值的信息具有重要的理论及应用价值. 近几年, 数据呈爆炸式增长, 数据的增长所引发的模式变革使机器学习和数据挖掘面临新的挑战, 数据约简是应对这种挑战的一种有效途径, 是机器学习和数据挖掘的重要环节.

机器学习使用实例数据和过去的经验训练学习算法, 以优化某种性能指标. 监督学习是机器学习的重要分支之一, 其目标是学习从输入到输出之间的映射关系, 其中输出的正确值已经给出. 在分类框架下, 监督学习使用有类标的数据训练学习算法, 主要任务是从有类标的数据中抽取规则, 以完成对新数据类标的预测, 是知识发现的一种重要手段. 从近几年的文献可以看出, 关于监督学习研究始终处于蓬勃发展的阶段, 研究人员提出许多监督学习的新方法和新技术, 并且成功应用于模式识别、故障诊断、生物信息处理、预测预报等许多领域.

本书第 1 章介绍了后续章节将要用到的预备知识, 包括分类与回归的概念、不确定性度量、数据约简的概念等. 第 2 章介绍了粗糙集及其扩展模型, 包括经典粗糙集模型、变精度粗糙集模型、相容粗糙集模型、粗糙模糊集模型和模糊粗糙集模型. 第 3 章介绍了求解分类问题的常用方法, 包括决策树、模糊决策树、支持向量机、极限学习机和概率神经网络. 第 4 章介绍了样例约简, 包括样例选择准则、交叉选择样例算法、压缩模糊 K 近邻规则、概率神经网络样例选择算法. 第 5 章介绍了属性约简, 包括特征子集评价准则、最小相关性最大依赖度属性约简、模糊属性约简及属性约简在模型选择中的应用.

本书的特点是结合作者近年来关于数据约简的研究成果, 在监督学习的框架下, 以分类问题为切入点, 从两个视角全面讨论了数据约简中的重要问题, 包括样例不确定性和重要性的度量、属性不确定性和重要性的度量、不确定性与分类系统泛化能力之间的关系、样例选择、特征选择、模型选择等问题. 本书大部分内容取材于作者的博士论文和近几年作者及其研究团队的相关研究成果. 借此机会, 特别感谢我的导师王熙照教授给予我的指导和帮助. 另外, 感谢我的研究生白晨燕、王华超、高原原、康晓萌、王婷婷、李胜杰、许宏雨、万丽艳、李塔、邵庆言和苗青对本书作出的贡献. 感谢研究生王敬庚、张垚、胡文祥、侯少星、王陈希, 他们对书稿

进行了校对. 本书得到了河北省自然科学项目: 粗糙集属性约简集成及其应用研究 (F2013201220), 河北省高等学校科学技术研究重点项目: 基于极端学习机的非平衡大数据分类研究 (ZD20131028) 和河北大学 “计算机应用技术” 省级重点学科的资助, 在此也表示感谢. 最后, 感谢科学出版社魏英杰老师的帮助.

限于作者水平, 书中的不足在所难免, 敬请各位同仁批评指正.

2015 年 1 月于河北大学

目　录

第1章　预 备 知 识

本章介绍后续章节将要用到的基础知识, 包括分类与回归的概念、随机变量不确定性度量、模糊集、样例约简和属性约简的形式化定义.

1.1　分类问题与回归问题

下面通过一个例子介绍什么是分类问题[1], 并在此基础上给出回归问题[2] 的定义. 因为本书在分类的框架下讨论问题, 所以本节重点介绍分类问题.

例 1.1.1　疾病诊断问题　设某疾病的诊断要化验 d 个指标 $a_1, a_2, \cdots, a_d$. 这些指标也称为属性或特征, 表 1.1 给出了 n 个患者的化验结果及医生的最终诊断结果. 其中, y_i 要么等于 $+1$, 要么等于 -1, $i = 1, 2, \cdots, n$. $y_i = +1$ 表示患有这种疾病, $y_i = -1$ 表示没有患这种疾病. 我们希望根据这些数据, 对新来的病人只检测这 d 个指标, 就可以推断该病人是否患有这种疾病, 这类问题就称为分类问题.

表 1.1　疾病诊断问题数据集

$\boldsymbol{x}$	a_1	a_2	$\cdots$	a_d	y
$\boldsymbol{x}_1$	x_{11}	x_{12}	$\cdots$	x_{1d}	y_1
$\boldsymbol{x}_2$	x_{21}	x_{22}	$\cdots$	x_{2d}	y_2
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
$\boldsymbol{x}_n$	x_{n1}	x_{n2}	$\cdots$	x_{nd}	y_n

如表 1.1 所示的数据集称为分类数据集, 也称为决策表, 可以用以下两种形式抽象地表示.

1) 用二元组表示

表 1.1 所示的分类数据集, 可用二元组 $(\boldsymbol{x}_i, y_i)$ 抽象地表示成如下形式, 即

$$D = \{(\boldsymbol{x}_i, y_i) | \boldsymbol{x}_i \in U, y_i \in C\} \tag{1.1}$$

其中, $\boldsymbol{x}_i$ 表示第 i 个样例; y_i 表示样例 $\boldsymbol{x}_i$ 所对应的类别标号, $i = 1, 2, \cdots, n$; $C = \{+1, -1\}$.

2) 用四元组表示

表 1.1 所示的分类数据集, 也可以抽象地表示为四元组, 即

$$DT = (U, A \cup C, V, f) \tag{1.2}$$

其中, $U = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n\}$ 是 n 个样例的集合; $A = \{a_1, a_2, \cdots, a_d\}$ 是 d 个描述对象 (或样例) 的条件属性 (或特征) 集合; C 是决策属性 (或类别属性) 集合; $V = V_1 \times V_2 \times \cdots \times V_d$ 是 d 个属性值域的笛卡儿积, V_i 是属性 a_i 的值域, $i = 1, 2, \cdots, d$; f 是信息函数 : $U \times A \to V$.

用式 (1.2) 表示的四元组也称为决策表, 为了描述方便, 本书中这两种等价表示会交替使用.

表 1.1 所述的分类问题是一个二类分类问题, 对于多类问题, y 的取值范围不再是 $\{+1, -1\}$, 而是由多个离散值构成的集合 $\{y_1, y_2, \cdots, y_k\}$, 如对于手写数字识别问题, y 的取值范围是 $\{0, 1, \cdots, 9\}$, 当然也可以用其他符号来表示, 如 $\{\omega_0, \omega_1, \cdots, \omega_9\}$. 下面针对多类分类问题, 从数学的角度给出分类的定义.

定义 1.1.1　给定分类数据集 $D = \{(\boldsymbol{x}_i, y_i) | \boldsymbol{x}_i \in U, y_i \in C\}$, 如果存在一个映射 $f : U \to C$, 使得对于任意的 $\boldsymbol{x}_i \in U$, 都有 $y_i = f(\boldsymbol{x}_i)$ 成立. 根据给定的分类数据集 D 寻找函数 $y = f(\boldsymbol{x})$ 的问题, 称为分类问题. 函数 $y = f(\boldsymbol{x})$ 也称为分类函数.

说明:

① 在分类问题中, 因变量 y 的取值范围是一个由有限个离散值构成的集合 C, 它相当于高级程序设计语言 (如 C++ 语言) 中的枚举类型. 若 C 变为实数集 $\mathbf{R}$ 或 $\mathbf{R}$ 中的一个区间 $[a, b]$, 则这类问题称为回归问题. 显然, 分类问题是回归问题的特殊情况.

② 函数 $y = f(\boldsymbol{x})$ 不一定有解析表达式, 可以用其他的形式, 如树、图或网络来表示.

③ 如果所有的 V_i 都是实数集 $\mathbf{R}$, 此时 $V = \mathbf{R}^d$.

下面举几个分类问题的例子.

例 1.1.2　天气分类问题　天气分类问题[3] 是一个两类分类问题, 用来预测什么样的天气条件适宜打网球. 天气数据集是机器学习领域中的一个经典数据集, 是包含 14 个样例的一个小数据集, 如表 1.2 所示.

表 1.2 天气分类问题数据集

$\boldsymbol{x}$	Outlook	Temperature	Humidity	Wind	y(PlayTennis)
$\boldsymbol{x}_1$	Sunny	Hot	High	Weak	No
$\boldsymbol{x}_2$	Sunny	Hot	High	Strong	No
$\boldsymbol{x}_3$	Cloudy	Hot	High	Weak	Yes
$\boldsymbol{x}_4$	Rain	Mild	High	Weak	Yes
$\boldsymbol{x}_5$	Rain	Cool	Normal	Weak	Yes
$\boldsymbol{x}_6$	Rain	Cool	Normal	Strong	No
$\boldsymbol{x}_7$	Cloudy	Cool	Normal	Strong	Yes
$\boldsymbol{x}_8$	Sunny	Mild	High	Weak	No
$\boldsymbol{x}_9$	Sunny	Cool	Normal	Weak	Yes
$\boldsymbol{x}_{10}$	Rain	Mild	Normal	Weak	Yes
$\boldsymbol{x}_{11}$	Sunny	Mild	Normal	Strong	Yes
$\boldsymbol{x}_{12}$	Cloudy	Mild	High	Strong	Yes
$\boldsymbol{x}_{13}$	Cloudy	Hot	Normal	Weak	Yes
$\boldsymbol{x}_{14}$	Rain	Mild	High	Strong	No

天气分类问题数据集有 14 个样例, 即 $U = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_{14}\}$; 4 个条件属性, 即 $A = \{a_1, a_2, a_3, a_4\}$, 其中, a_1 = Outlook, a_2 = Temperature, a_3 = Humidity, a_4 = Wind, 它们都是离散值属性, 相当于高级程序设计语言中的枚举类型属性. $V = V_1 \times V_2 \times V_3 \times V_4$, $V_1 = \{\text{Sunny}, \text{Cloudy}, \text{Rain}\}$, $V_2 = \{\text{Hot}, \text{Mild}, \text{Cool}\}$, $V_3 = \{\text{High}, \text{Normal}\}$, $V_4 = \{\text{Strong}, \text{Weak}\}$. 决策属性集合由单决策属性构成, 即 $C = \{y\}$, y = PlayTennis, 它只取 Yes 和 No 两个值, 所以天气分类问题是一个两类分类问题. 显然, 从该数据集中找到的分类函数 $y = f(\boldsymbol{x})$ 不可能有解析表达式. 在第 3 章, 我们将会看到 $y = f(\boldsymbol{x})$ 可用一棵树来表示.

例 1.1.3 鸢尾花分类问题 鸢尾花分类问题是一个三类分类问题, 它根据花萼长 (Sepal length)、花萼宽 (Sepal width)、花瓣长 (Petal length) 和花瓣宽 (Petal width) 四个条件属性对鸢尾花进行分类. 鸢尾花数据集 Iris[4] 包含三类 150 个样例, 每类 50 个样例, 如表 1.3 所示.

Iris 数据集有 150 个样例, 即 $U = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_{150}\}$; 4 个条件属性, 即 $A = \{a_1, a_2, a_3, a_4\}$, 其中, a_1 = Sepal length, a_2 = Sepal width, a_3 = Petal length, a_4 = Petal width, 它们都是连续值属性. $V = V_1 \times V_2 \times V_3 \times V_4$, $V_1 = V_2 = V_3 = V_4 = R$, 即 $V = \mathbf{R}^4$. 决策属性集合由单决策属性构成, 即 $C = \{y\}$,

$y \in \{\text{Iris-setosa}, \text{Iris-versicolor}, \text{Iris-virginica}\}$. 由于 Iris 数据集中四个条件属性都是连续值属性, 所以该数据集是一个连续值数据集.

表 1.3　鸢尾花分类问题数据集

$\boldsymbol{x}$	a_1	a_2	a_3	a_4	y
$\boldsymbol{x}_1$	5.1	3.5	1.4	0.2	Iris-setosa
$\boldsymbol{x}_2$	4.9	3.0	1.4	0.2	Iris-setosa
⋮	⋮	⋮	⋮	⋮	⋮
$\boldsymbol{x}_{50}$	5.0	3.3	1.4	0.2	Iris-setosa
$\boldsymbol{x}_{51}$	7.0	3.2	4.7	1.4	Iris-versicolor
$\boldsymbol{x}_{52}$	6.4	3.2	4.5	1.5	Iris-versicolor
⋮	⋮	⋮	⋮	⋮	⋮
$\boldsymbol{x}_{100}$	5.7	2.8	4.1	1.3	Iris-versicolor
$\boldsymbol{x}_{101}$	6.3	3.3	6.0	2.5	Iris-virginica
$\boldsymbol{x}_{102}$	5.8	2.7	5.1	1.9	Iris-virginica
⋮	⋮	⋮	⋮	⋮	⋮
$\boldsymbol{x}_{150}$	5.9	3.0	5.1	1.8	Iris-virginica

例 1.1.4　助教评估分类问题　助教评估分类问题也是一个三类分类问题, 它根据母语是否是英语 (A native English speaker)、课程讲师 (Course instructor)、课程 (Course)、是否正常学期 (A regular semester) 和班级规模 (Class size) 五个条件属性对助教评估分类. 助教评估分类数据集 (Teaching Assistant Evaluation, TAE)[4] 包含三类 151 个样例, 第一类 (Low)49 个样例, 第二类 (Medium)50 个样例, 第三类 (High)52 个样例, 如表 1.4 所示.

表 1.4　助教评估分类问题数据集

$\boldsymbol{x}$	a_1	a_2	a_3	a_4	a_5	y
$\boldsymbol{x}_1$	2	21	2	2	42	Low
$\boldsymbol{x}_2$	2	22	3	2	28	Low
⋮	⋮	⋮	⋮	⋮	⋮	⋮
$\boldsymbol{x}_{49}$	2	2	10	2	27	Low
$\boldsymbol{x}_{50}$	2	6	17	2	42	Medium
$\boldsymbol{x}_{51}$	2	6	17	2	43	Medium
⋮	⋮	⋮	⋮	⋮	⋮	⋮
$\boldsymbol{x}_{99}$	2	22	1	2	42	Medium
$\boldsymbol{x}_{100}$	1	23	3	1	19	High
$\boldsymbol{x}_{101}$	2	15	3	1	17	High
⋮	⋮	⋮	⋮	⋮	⋮	⋮
$\boldsymbol{x}_{151}$	2	20	2	2	45	High

TAE 数据集有 151 个样例, 即 $U = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_{151}\}$; 5 个条件属性, 即 $A = \{a_1, a_2, \cdots, a_5\}$, 其中, $a_1 = A$ native English speaker, $a_2 =$ Course instructor, $a_3 =$ Course, $a_4 =$ A regular semester, $a_5 =$ Class size. 其中, a_1 表示母语是否是英语, 是一个二值属性; a_2 表示课程讲师, 共 25 个课程讲师, 每个课程讲师用一个符号值表示, 共 25 个值; a_3 表示助教课程, 共 26 门课程, 每门课程用一个符号值表示, 共 26 个值; a_4 表示是否正常学期, 是一个二值属性; a_5 表示班级规模, 是一个数值属性. 显然, TAE 数据集是一个混合类型数据集.

1.2 不确定性度量

不确定性在机器学习中是一种常见的现象, 存在于学习过程的各个环节, 如数据预处理 (包括特征选择和样例选择)、算法设计、模型选择等, 它对学习系统的性能有重要的影响. 常见的不确定性包括随机性、模糊性和粗糙性. 随机性[5] 是客观存在的一种不确定性. 模糊性[6] 是人类在认识客观实际的过程中, 由于无法给出清晰准确的界限而产生的一种不确定性, 是一种认知不确定性. 粗糙性[7] 是由于人类掌握的知识不充分而产生的一种不确定性, 是一种知识不确定性. 本节介绍前两种不确定性的度量, 粗糙性度量在第 2 章详细介绍.

1.2.1 随机变量的不确定性度量

熵是随机变量不确定性的度量, 下面分两种情况给出熵的定义, 并讨论其性质.

1. 离散型随机变量不确定性度量

1) 熵

设 X 是离散型随机变量, 它所有可能取值的集合为 $\mathcal{X}$, 对于任意的 $x \in \mathcal{X}$, 令 $\Pr\{X = x\} = p(x)$, X 服从的概率分布为 $p(x)$, 记为 $X \sim p(x)$, 下面给出熵的定义[8].

定义 1.2.1 离散型随机变量 X 的熵定义为

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log_2 p(x) \tag{1.3}$$

说明:

① 熵的单位为 bit, 当公式 (1.3) 中的对数变成以 e 为底的对数, 即自然对数时, 熵的单位为 net.

② 随机变量 X 的熵表示它取值的混乱程度, 即不确定性程度.

③ 随机变量 X 的熵也可以写成 $H(p)$.

④ 熵是随机变量 X 的分布函数, 不依赖于 X 的具体取值, 而依赖于取值的概率.

设 E 是期望算子, 如果 $X \sim p(x)$, 则随机变量 X 的函数 $g(X)$ 的期望值为

$$E_p g(X) = \sum_{x \in \mathcal{X}} g(x)p(x) \tag{1.4}$$

当 $g(X) = \log_2 \dfrac{1}{p(X)}$, 则 X 的熵有如下定义形式, 即

$$H(X) = E_p \log_2 \frac{1}{p(X)} \tag{1.5}$$

因为 $0 \leqslant p(x) \leqslant 1$, 所以 $\log_2 \dfrac{1}{p(X)} \geqslant 0$, 从而有 $H(X) \geqslant 0$.

例 1.2.1　　设 $\mathcal{X} = \{0, 1\}$, 且 $\Pr(X = 1) = p$, 求随机变量 X 的熵.

因为 X 服从 0-1 分布, 所以 $\Pr(X = 0) = 1 - p$. 根据式 (1.3), 随机变量 X 的熵为

$$H(X) = -p \times \log_2 p - (1 - p) \times \log_2(1 - p) \tag{1.6}$$

从式 (1.6) 可以看出, 随机变量 X 的熵是 p 的函数 $H(p)$. 当 $p = \dfrac{1}{2}$ 时, 熵的值最大, 等于 1. $H(p)$ 的图形如图 1.1 所示.

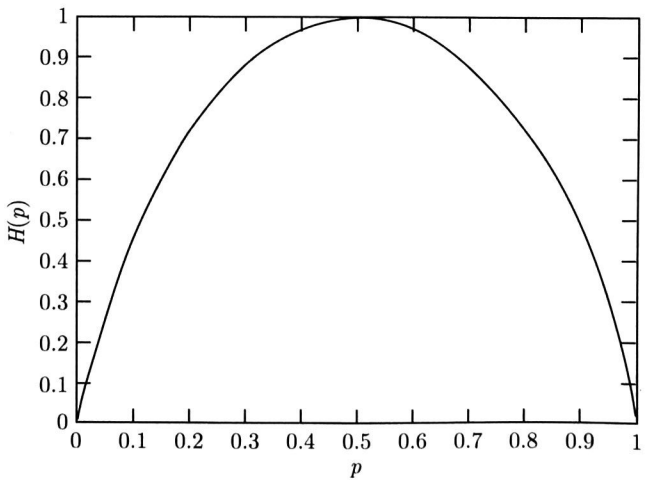

图 1.1　概率分布 p 与熵的关系

例 1.2.2 求表 1.2 所示天气数据集中 4 个条件属性和 1 个决策属性的熵.

先求第一个条件属性 Outlook 的熵. 此时, $X=\text{Outlook}$, $\mathcal{X}=\{\text{Sunny},\text{Cloudy},\text{Rain}\}$. 由表 1.2 可以求出: $\Pr(X=\text{Sunny})=p_1=\dfrac{5}{14}$, $\Pr(X=\text{Cloudy})=p_2=\dfrac{4}{14}$, $\Pr(X=\text{Rain})=p_3=\dfrac{5}{14}$. 根据式 (1.3), 随机变量 X 的熵, 即第一个条件属性 Outlook 的熵为

$$H(\text{Outlook})=-\frac{5}{14}\log_2\frac{5}{14}-\frac{4}{14}\log_2\frac{4}{14}-\frac{5}{14}\log_2\frac{5}{14}=1.58$$

同理, 可得其他 3 个条件属性的熵分别为

$$H(\text{Temperature})=-\frac{4}{14}\log_2\frac{4}{14}-\frac{6}{14}\log_2\frac{6}{14}-\frac{4}{14}\log_2\frac{4}{14}=1.56$$

$$H(\text{Humidity})=-\frac{7}{14}\log_2\frac{7}{14}-\frac{7}{14}\log_2\frac{7}{14}=1.00$$

$$H(\text{Wind})=-\frac{8}{14}\log_2\frac{8}{14}-\frac{6}{14}\log_2\frac{6}{14}=0.99$$

决策属性 $y=\text{PlayTennis}$ 的熵为

$$H(\text{PlayTennis})=-\frac{5}{14}\log_2\frac{5}{14}-\frac{9}{14}\log_2\frac{9}{14}=0.94$$

2) 联合熵与条件熵

上面定义了单个离散型随机变量的熵, 现在将这一定义推广到离散型随机变量对 (X,Y) 上去.

定义 1.2.2 设离散型随机变量对 (X,Y) 服从的联合分布为 $p(x,y)$, 即 $(X,Y)\sim p(x,y)$,(X,Y) 的联合熵定义为

$$H(X,Y)=-\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}p(x,y)\log_2 p(x,y) \tag{1.7}$$

类似地, (X,Y) 联合熵的定义还可以表示为

$$H(X,Y)=-E\log_2 p(X,Y) \tag{1.8}$$

下面给出条件熵的定义.

定义 1.2.3 设 $(X,Y)\sim p(x,y)$, 在给定 X 的条件下, Y 的条件熵定义为

$$\begin{aligned}H(Y|X)&=-\sum_{x\in\mathcal{X}}p(x)H(Y|X=x)\\&=-\sum_{x\in\mathcal{X}}p(x)\sum_{y\in\mathcal{Y}}p(y|x)\log_2 p(y|x)\end{aligned}$$

$$\begin{aligned} &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log_2 p(y|x) \\ &= -E\log_2 p(Y|X) \end{aligned} \tag{1.9}$$

定理 1.2.1　(链规则)

$$H(X,Y) = H(X) + H(Y|X) \tag{1.10}$$

证明:

$$\begin{aligned} H(X,Y) &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log_2 p(x,y) \\ &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log_2 p(x)p(y|x) \\ &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log_2 p(x) - \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log_2 p(y|x) \\ &= -\sum_{x\in\mathcal{X}} p(x)\log_2 p(x) - \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log_2 p(y|x) \\ &= H(X) + H(Y|X) \end{aligned} \tag{1.11}$$

■

推论 1.2.1　设 X,Y,Z 是三个离散型随机变量, 根据定理 (1.2.1), 可得如下推论

$$H(X,Y|Z) = H(X|Z) + H(Y|X,Z) \tag{1.12}$$

说明:

一般地, $H(Y|X) \neq H(X|Y)$, 但有 $H(X) - H(X|Y) = H(Y) - H(Y|X)$.

3) 相对熵与互信息

一个随机变量的熵是其不确定性的度量, 而相对熵也称为 K-L 散度, 是一个随机变量的两个概率分布之间距离的度量. 相对熵用符号 $D(p||q)$ 表示, 从统计的角度[5,8], $D(p||q)$ 描述的是命题: "假设分布是 q, 但真正的分布是 p" 的无效性度量.

下面给出相对熵的定义.

定义 1.2.4　一个随机变量 X 的两个概率分布 $p(x)$ 和 $q(x)$ 之间的相对熵定义为

$$\begin{aligned} D(p||q) &= \sum_{x\in\mathcal{X}} p(x)\log_2\frac{p(x)}{q(x)} \\ &= E_p\log_2\frac{p(X)}{q(X)} \end{aligned} \tag{1.13}$$