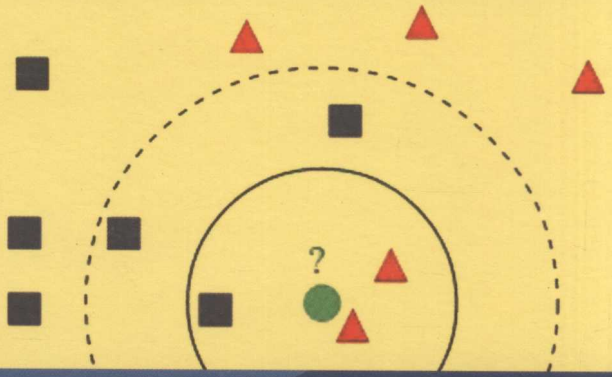


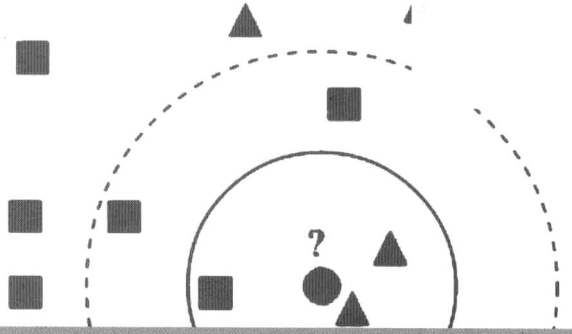
近邻分类方法及其应用

[上册]

郭躬德 陈黎飞 李南 ◎著

Nearest Neighbour Classification Method and its Applications

厦门大学出版社 XIAMEN UNIVERSITY PRESS
国家一级出版社
全国百佳图书出版单位



近邻分类方法及其应用

[上册]

郭躬德　陈黎飞　李南　◎著

厦门大学出版社 XIAMEN UNIVERSITY PRESS
国家一级出版社
全国百佳图书出版单位

图书在版编目(CIP)数据

近邻分类方法及其应用．上册/郭躬德，陈黎飞，李南著．—厦门：厦门大学出版社，2013.12
ISBN 978-7-5615-4856-1

Ⅰ.①近… Ⅱ.①郭…②陈…③李… Ⅲ.①数据采掘-算法分析 Ⅳ.①TP311.131

中国版本图书馆 CIP 数据核字(2013)第 281153 号

厦门大学出版社出版发行
(地址：厦门市软件园二期望海路 39 号 邮编：361008)
http://www.xmupress.com
xmup @ xmupress.com
厦门市明亮彩印有限公司印刷
2013 年 12 月第 1 版 2013 年 12 月第 1 次印刷
开本：787×1092 1/16 印张：9.5
插页：2 字数：218 千字
定价：26.00 元
本书如有印装质量问题请直接寄承印厂调换

前 言

计算机技术的普及应用给人类社会带来了深刻的变革，也使得我们所拥有的数据以前所未有的速度膨胀。随着大数据时代的到来，越来越多的人开始关注数据挖掘这一项大数据分析和处理的重要技术。作为数据挖掘的一种主要方法，分类(classification)——用于发掘隐藏在历史数据中的类别模式进而对未知事件做出预测或判断的技术——由于其很强的实用性成为了许多数据分析处理系统的基本构件。在机器学习领域，分类是有监督学习的代表性方法，其应用也已深入到信息检索、生物信息、客户关系管理等社会经济生活的方方面面。

近邻分类技术源来已久，可以追溯到早期对 k_n-NN 规则(E. Fix 等人，1951 年)和 k-NN 规则(T. Cover 等人，1967 年)的研究。随后发展起来的 kNN 分类算法由于其原理简单、易于实现、可扩展性好、可解释性强等优点，备受青睐，位列 2005 年 ICDM 国际会议遴选的 10 大最有影响力的数据挖掘算法之一。如今，近邻分类方法无论在理论模型、算法还是在领域应用方面都吸引了众多的研究者和实践者，呈现出蓬勃发展的态势，涌现了一大批改进型新算法，也提出了一些基于近邻分类思想的新模型、新方法。“欲穷千里目，更上一层楼”，跟踪了解近年来取得的这些新进展才能进一步推进近邻分类技术的研究和深入应用，这也是本书出版的首要目的所在。

著者之一郭躬德教授系英国 Ulster 大学归国学者，长期从事近邻分类方法的研究与应用，早于 2003 年提出称为 kNNModel 的近邻分类新方法，赢得业界热烈反响。近年来，在郭教授的带领下，福建师范大学数据挖掘与网络内容安全实验室开展了近邻分类理论方法与应用方面的系统研究，取得了一系列成果。在理论方法方面，研究团队提出了基于近邻思想的相似性度量新方法并将之推广到类属型数据，提出了增量学习、多代表点学习和

子空间近邻分类等新方法；应用研究涵盖了毒性物质预测、特征选择、文本分类以及数据流分类等近邻分类的新应用领域。本书将有关研究成果集结成册，以飨读者。

本书共六章，第一章介绍近邻分类算法及近年来的研究新进展，第二章至第六章每章介绍近邻分类的一类新方法。郭躬德主要编写本书的第二、三章，并参与编写第四、六章部分章节，约 20 万字；陈黎飞主要编写第四、五、六章，约 30 万字；李南主要编写第一章，并参与编写第四章部分章节，约 5 万字。研究生黄彧、陈红、辛轶、黄杰、李南、张健飞、卢伟胜、兰天，访问学者陈雪云等参加了有关研究工作和部分章节写作。在写作过程中，参考了大量的国内外文献资料，在此一并表示感谢。

本书内容有误或不妥之处，欢迎读者批评指正。

郭躬德　陈黎飞　李南
福建师范大学数学与计算机科学学院
2014 年 1 月

目 录

上 册

下 册

第1章　近邻分类方法及其演变

1.1 分类概念、算法

【摘要】本节主要介绍数据挖掘的主要任务，分类的概念、过程和分类模型评价，以及常用的分类算法及其应用。

1.1.1 数据挖掘

伴随着信息技术的高速发展，各行各业积累的数据量也急剧增长，人们迫切地需要将这些数据转化为有用的知识以促进生产力的发展。因此，能够满足人们此类需求的数据挖掘技术逐渐受到了信息产业界乃至整个社会的关注。

数据挖掘一般通过分析给定数据集里的数据来解决问题。例如，在激烈的市场竞争中，如何牢牢地把握住客户一直是商家关注的首要问题。一个关于客户当前所选择的业务以及其个人资料的数据库是解决这个问题的关键。通过分析客户以往的行为模式，挑选出那些对即将提出的新业务最有可能感兴趣的客户群体，进而向他们提供特殊建议以推广该项新业务。值得注意的是，对整个客户群体进行业务推广的代价是高昂的，使用数据挖掘技术选择特定的群体能够有效地节约成本。概括地说，数据挖掘的目的就是从海量数据中提取隐含在其中的、事先未知但又潜在有用的知识和可以理解的模式，进而帮助人们做出更为准确、客观的分析和判断[1]。数据挖掘的主要任务包括关联规则发现[2]、聚类分析[3]、分类及回归[4]等。

1. 关联规则发现

关联规则的作用是从大量数据中挖掘出有价值的数据项之间的关系。经典的实例是购物篮分析(market basket analysis)。超市对关于顾客购买记录的数据库进行关联规则挖掘，从而发现顾客的购买习惯。一个有趣的例子是沃尔玛超市通过数据挖掘中的关联规则挖掘，惊奇地发现顾客在购买啤酒的同时，经常也会一起购买尿布(这种独特的销售现象出现在年轻的父亲身上)。于是，超市就调整货架的布局，将啤酒和尿布放在一起以增进销量。

2. 聚类分析

聚类是把数据样本集合分成由相似对象组成的多个子集(簇)的过程,使得在同一个簇中的样本具有相似的一些属性。聚类的一个商业应用是电子商务中的网站建设。具体的步骤是:首先使用聚类算法找出具有相似浏览行为的用户,然后分析这些用户的共同特征。这些共同特征能够帮助商家更好地了解自己的客户,进而向客户提供更合适的服务。

3. 分类及回归

分类(classification)和回归(regression)是两种不同的数据分析形式,都可以用来建立预测未来数据的模型。分类用于预测对象的离散类别,回归则用于预测对象的连续或者有序取值。分类主要用于定性,例如基于历史数据建立分类模型,预测明天的股市大盘指数是涨还是跌。而回归则用于定量,对应的就是建立好的模型预测明天的大盘指数是涨多少点还是跌多少点。本书主要关注于数据挖掘中的分类问题,关于分类的详细内容将在后续章节做具体介绍。

按照是否使用数据样本的类别属性来划分,数据挖掘的学习方法可以分为监督学习(supervised learning)、半监督学习(semi-supervised learning)和无监督学习(unsupervised learning)三种[5]。其中,监督学习利用样本的类别信息来指导学习过程,无监督学习在学习中使用的数据均是未标记类别的,而半监督学习主要考虑如何同时利用大量没有类别标记的数据和少量已标记类别的数据进行学习。

1.1.2 分类的概念

1.1.2.1 分类的过程

作为数据挖掘领域的一个重要分支,众多学者对分类问题[6]进行了深入研究。承上所述,区别于回归方法,分类的输出是离散的类别值,而回归方法则是连续或有序值。分类的基本过程如图 1-1 所示。由图 1-1 可以看出,分类过程依次经历了 5 个不同的阶段,以下将对每个阶段进行详细介绍。

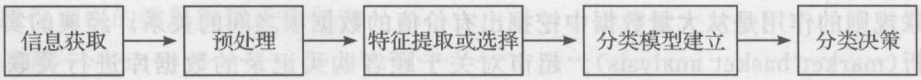


图 1-1 分类的过程

1. 信息获取

信息获取是对研究对象进行测量和量化，使其变成计算机可以处理的形式(例如矩阵或向量)。通常情况下，通过信息获取的数据集中的每个样本可以由 $d+1$ 个属性描述的数据库元组来表示(一个数据库元组对应于一个向量)，形如：

$$X=\{x_1,x_2,\cdots,x_d,y\}$$
$$s.t.\ d>0$$

其中 x_i 表示样本 X 第 i 维的属性值，y 表示类标号属性，即样本 X 的类别。由于事先知道数据集中每个样本的类别，因此分类属于监督学习的一种。其中，样本属性的类型可以是：

(1)数值型(numeric)：可以是整数或者实数，比如人的体重和年龄。

(2)命名型(nominal)：例如某个描述天气数据集中的湿度(humidity)属性。该属性共有两种取值：高(high)和普通(normal)。数据集中每个样本对应的“humidity”值必是这二者之一。命名型属性的一种特殊形式是类属型(categorical)属性。例如描述一个人性别的属性，此属性的两种取值(“男”和“女”)之间没有顺序关系。

(3)字符串型(string)：可以包含任意的文本，在文本分类中非常有用。

(4)日期和时间型(date)等。

表 1-1 是一个关于特定天气下能否进行某项活动的数据集。数据集中样本的 3 个命名型属性分别是阴晴(outlook，取值范围{sunny，overcast，rainy})、湿度(humidity，取值范围{high，normal})以及刮风(windy，取值范围{true，false})。这里的 outlook 可以看作是类属型属性。样本还有 1 个数值型属性温度(temperature)，而玩(play)是类标号属性。表中的每一行对应于数据集中的一条样本。

表 1-1　天气数据集

outlook	temperature	humidity	windy	class：play
sunny	85	high	false	no
sunny	80	high	true	no
overcast	83	high	false	yes
rainy	70	normal	false	yes
overcast	81	normal	false	yes
…	…	…	…	…

2. 预处理

在分类之前，为了提高所建立分类模型的有效性，通常要对数据进行预处理操作。常用

的预处理过程包括数据清理、数据变换等。

(1)数据清理

现实中可能会发生测量设备故障、人工录入错误等情况,因此实际使用的数据集中不可避免地会出现收集到的数据不完整或者包含一些错误等问题。数据清理包括填充样本的缺失属性值、减少或者消除噪声数据以及清除重复数据等操作。

对于数据集中数据不完整问题,常用的处理方法有[7]:

方法 1(常量替代法):所有数据中缺失属性的取值都用同一个常量来填充,比如"Error"。该方法最为简单,但是它并不十分可靠。

方法 2(平均值替代法):采用数据集上某属性的平均值代替不完整样本该属性上的缺失值。

方法 3(最常见值替代法):使用同一属性中出现最多的取值作为不完整样本该属性上的缺失值。

方法 4(估算值替代法):采用相关算法(例如回归算法等)预测不完整样本缺失属性的可能值。

所谓的噪声数据,指的是属性值中存在随机错误或者偏差的样本。例如,如果数据集中一条样本的"年龄"属性值是"-10",这显然不符合常理,那么这条样本就属于噪声样本。通俗地说,噪声样本就是"错误样本"。基于噪声数据建立的模型的分类效果往往不尽如人意。消除噪声数据的方法除了人工检查校对之外,还可以使用聚类分析技术。具体的步骤是先将原始数据集分成固定数目的子集(簇),那些处于较小簇中(簇内样本的数目远远小于其他簇)的样本往往就是噪声样本,通过删除这些样本来达到提高模型有效性的目的。

(2)数据变换

数据变换的目的是将数据转换或者统一为某种适合数据挖掘的形式,常用的做法包括规范化、离散化等。

规范化通常是为了消除不同属性取值区间的差异对相似性度量带来的影响,因而将数据变换到指定区间(如[0,1])内。常用的规范化方法包括:

方法 1:最小—最大规范化

设 $\min_i$ 和 $\max_i$ 分别为数据集中样本第 i 个属性的最小值和最大值。最小—最大规范化通过计算

$$x_i'=\frac{x_i-\min_i}{\max_i-\min_i}(new\_\max_i-new\_\min_i)+new\_\min_i$$

将第 i 个属性的取值 x_i 映射到区间[$new\_\min_i$,$new\_\max_i$]中的 x_i'。

设数据集中样本第 i 个属性的取值分别为最大值 $\max_i=50$,最小值 $\min_i=1$。如果使用最小—最大规范化,将某样本该属性的取值 $x_i=30$ 映射到[0,1],则有:

$$x_i'=\frac{30-1}{50-1}(1-0)+0=\frac{29}{49}$$

方法 2：Z-score 规范化

将数据集中样本的数目用 n 来表示，所有样本第 i 个属性上的取值以 $\{x_{i1},x_{i2},\cdots,x_{in}\}$ 表示，$\overline{x}$ 和 σ 分别为 $\{x_{i1},x_{i2},\cdots,x_{in}\}$ 的平均值和标准差，即：

$$\overline{x}=\frac{\sum_{j=1}^{n}x_{ij}}{n}\qquad \sigma=\sqrt{\frac{\sum_{j=1}^{n}(x_{ij}-\overline{x})^2}{n}}$$

那么规范化之后

$$x_{ij}'=\frac{x_{ij}-\overline{x}}{\sigma}$$

设数据集中样本的第 i 个属性的平均值 $\overline{x}=54000$，标准差 $\sigma=16000$。如果使用 Z-score 规范化，那么某样本该属性的取值 $x_i=73600$ 则转换为：

$$x_i'=\frac{73600-54000}{16000}=1.225$$

多数分类算法只能处理某种特定类型的属性。**离散化**的作用就是将数据集中样本的数值型属性转换为命名型属性。离散化时可以对数据先使用等宽或等频分箱，然后用箱均值、中位数或者某个固定值替换箱中的每个值，就能够将属性值离散化。以某个数据集中的“年龄”属性为例子，因为属性取值大于 0，于是将其分成几个不同的区间：[0,29]、[30,45]、[46,+∞)，原始数据集中该属性上对应的取值分别用 young、mid-age 以及 old 来替代。例如，如果原始数据集中某条样本的“年龄”属性的取值是 27，那么经过上述离散化之后，该样本“年龄”属性的取值就相应离散化成了 young。

3. 特征提取或选择

由于在现实应用中，数据的属性个数往往较多，直接对这些原始数据进行分类往往会影响分类效率。特征提取或选择的作用就是抽取或通过变换得到最能反映每个类别本质的属性，进而利用这些属性代表原始数据中的所有属性进行分类。

特征提取[8]通常通过映射（或变换）的方法获取最有效的特征。经过映射后的特征称为二次特征，通常是原始特征的某种组合。传统的特征提取方法可以分为线性和非线性两类。目前主流的线性特征提取方法有主成分分析法（principle component analysis，简称 PCA）[9]、Fisher 线性鉴别分析法（Fisher linier discriminant analysis，简称 FLD）[10]等。通过求解样本协方差矩阵的特征值和特征向量，PCA 方法试图找到方差最大的特征。FLD 方法的目标是保证样本在新空间中有最大的类间距离和最小的类内距离，即样本在该空间中有最佳的可分离性。线性自组织映射（self-organizing feature map，简称 SOM）[11]是一

种非线性的特征提取方法,该方法的目标是利用低维空间中的样本点来表示原始高维空间中的样本点,使得低维空间的样本之间尽可能保持原始空间中的距离和相似性关系。特征提取主要应用于样本特征维度较高的面部表情识别、汉字以及英文字母识别、图像检索等领域。

特征选择是从全部特征中选取一个特征子集。特征选择的具体步骤[12]是:

Step1:从特征全集中产生出一个特征子集。

Step2:用评价函数对该特征子集进行评价。

Step3:将评价的结果与停止准则进行比较。若评价结果达到要求就停止,否则重复以上步骤。

Step4:验证所选取特征子集的有效性。

在 Step1 中使用的产生特征子空间算法可以分为完全搜索、启发式搜索以及随机搜索三种。广度优先搜索是一种最简单的完全搜索方法,属于穷举搜索,需要遍历所有组合以产生最优的特征子集。显然,如果特征空间维度较高,这种方法所需要的时间令人难以接受。启发式搜索的方法有序列前向选择、序列后向选择等。序列前向选择方法每次都选择一个使得评价函数取值达到最优的特征加入,其实就是一种简单的贪心算法。该方法的缺点是只能加入特征而不能去除特征。序列后向选择方法则恰好相反,每次删除一个特征。随机搜索的方法包括使用遗传算法、模拟退火算法等。

根据不同的工作原理,Step2 中的评价函数可以分为筛选器(filter)和分装器(wrapper)两类。前者通过分析特征子集内部的特点来衡量其好坏,与分类器的选择无关。例如可以使用距离作为评价标准,即好的特征子集应该使得来自同一类别的样本之间的距离尽可能小、来自不同类别的样本之间的距离尽可能大。而后者则使用指定的算法,利用选取的特征子集对样本集进行分类,模型的分类精度被用作衡量特征子集好坏的标准。

4. 分类模型建立

在建立分类模型之前,需要将给定的数据集随机地分为训练数据集和测试数据集两个部分(需要划分的原因将在 1.2.2 节中说明)。在分类模型建立阶段,通过分析训练数据集中属于每个类别的样本,使用分类算法建立一个模型对相应的类别进行概念描述。通过学习得到的分类模型的形式可以是分类规则、决策树等,主要的分类模型将在第 1.1.3 节中进行详细介绍。

在建立好分类模型之后,还需要在测试数据集上对分类模型的有效性进行测试,此时通常使用分类精度作为评价标准。对于测试数据集上的每一个样本,如果通过已经建立的分类模型预测出来的类别与其真实的类别相同,那么说明分类正确,否则,说明分类错误。如果测试数据集上所有样本的平均分类精度可以接受,那么在分类决策阶段就可以使用该模

型对未知类别的待分类样本进行类别预测。需要说明的是，之所以使用不同于训练数据集的样本作为测试数据集，是因为基于训练数据集所建立的分类模型对于自身样本的评估往往是乐观的，这并不能说明分类模型对未知样本的分类是有效的。

举一个著名的“鸢尾花分类”的例子。事先给定包含 150 条样本的鸢尾花 Iris 数据集（该数据集可以从 http://www.ics.uci.edu/～mlearn/databases/下载），它包含了三种鸢尾花种类——setosa、versicolor 和 virginica，每种各有 50 条样本。每条样本记录了一朵鸢尾花的花萼长（sepal length）、花萼宽（sepal width）、花瓣长（petal length）和花瓣宽（petal width）4 个属性以及该样本属于哪一种鸢尾花（即类别属性），具体见表 1-2 所示。

表 1-2　鸢尾花数据

sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	class
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
7.0	3.2	4.7	1.4	versicolor
6.3	3.3	6.0	2.5	virginica
6.3	2.9	5.6	1.8	virginica
…	…	…	…	…

首先按照 2∶1 的比例，将数据集随机地分为训练数据集和测试数据集两个部分。在训练数据集上，利用分类算法建立一个分类模型（具体的分类算法将在第 1.3.3 节中详细介绍），再使用测试数据集测试模型的分类精度。分类过程中模型建立的流程见图 1-2，其中分类模型使用形如：

$$\text{If} \quad petal\ width \leqslant 0.6 \quad \text{Then} \quad class = \text{setosa}$$

的分类规则来表示。

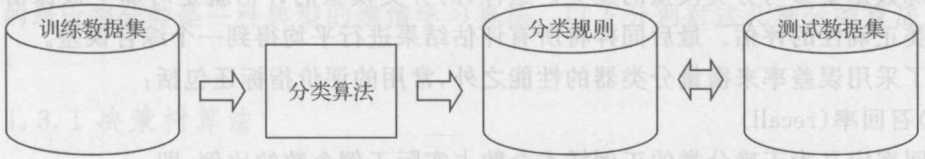


图 1-2　分类过程中的模型建立流程

5. 分类决策

如果所建立的模型的分类有效性可以接受，那么在输入未知类别样本的花萼长、花萼

宽、花瓣长和花瓣宽4个属性之后，就能够利用该分类模型对其类别进行预测，判断它属于哪一种类别的鸢尾花。例如，输入未知样本 $X'=\{sepal\ length=5.0, sepal\ width=3.6, petal\ length=1.4, petal\ width=0.3\}$，那么根据前述分类规则就将其分为 setosa 类，具体流程如图 1-3 所示。

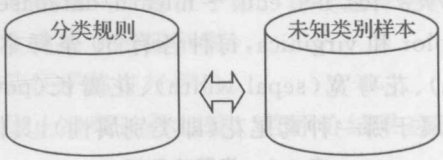


图 1-3　分类过程中的分类决策流程

1.1.2.2 分类模型评价

承上节所述，人们所关注的是分类模型对于未来新数据的分类效果，而非旧数据。给定数据集中每个样本的类别都是已知的，因此才能用它进行训练，但是通常对这些样本的分类并不感兴趣。因此，需要将给定数据集划分为训练数据集和测试数据集两个部分，以测试数据集上的分类结果来近似评估分类模型对未来新数据的分类效果。划分的方法通常有以下两种：

1. n 折交叉验证

使用 n 折(n-fold)交叉验证时，数据集被随机分为 n 个部分(通常 $n=10$)，每一部分中的类比例与整个数据集中的类比例基本一致。每个部分被轮流作为测试数据集，其余 $n-1$ 个部分作为训练数据集。这样一共进行了 n 次学习，每次使用不同的训练数据集。最后，将 n 次学习的误差进行平均得到一个综合误差。

2. 留一交叉验证

留一(leave-one-out)交叉验证依次将每个样本作为测试数据集，而剩下的所有样本则作为训练数据集参与分类模型的建立。这样，对分类模型的评估就是对那个被保留在外的样本分类正确性的评估。最后同样将所有评估结果进行平均得到一个综合误差。

除了采用误差率来衡量分类器的性能之外，常用的评价指标还包括：

(1)召回率(recall)

召回率定义为正确分类的正例样本个数占实际正例个数的比例，即：

$$recall=\frac{true\ positive}{(true\ positive)+(false\ negative)}$$

(2)查准率(precision)

查准率定义为正确分类的正例样本个数占分类为正例的样本个数的比例，即：

$$precision=\frac{true\ positive}{(true\ positive)+(false\ positive)}$$

在上述公式中，对于给定类别 c，*true positive* 表示测试样本中属于类 c 且被分类模型正确分类的样本数目，*false positive* 表示测试样本中不属于类 c 但是被分类模型错分到类 c 的样本数目，*false negative* 表示测试样本中属于类 c 但是被分类模型错分到其他类的样本数目。

(3) $F_1-measure$

$F_1-measure$ 是查准率(precision)以及召回率(recall)的调和平均，即：

$$F_1(recall,precison)=\frac{2\times recall\times precison}{recall+precision}$$

(4) $macroaveraged-F_1$

$macroaveraged-F_1$ 是每个类别 $F_1-measure$ 指标的算术平均值。设给定训练数据集有 m 类，$F_1(i)$ 为第 i 类的 $F_1-measure$ 值，那么

$$macroaveraged-F_1=\frac{\sum_{i=1}^{m}F_1(i)}{m}$$

其余的评价标准包括 ROC 曲线(receiver operating characteristic curve)与 AUC(the area under the ROC curve)[13]等。

1.1.3 分类算法

按照训练数据集的学习方式划分，分类算法可以分为急切(eager)型和懒惰(lazy)型两种[14]。急切学习算法在模型建立阶段以训练数据集为基础建立模型，然后将建立好的模型用于未知样本类别的预测。与其不同的是，懒惰学习算法在训练阶段不建立模型，而是直接利用训练样本对待分类样本的类别进行预测。学者们已经提出的急切学习算法包括决策树算法[15]、支持向量机(SVM)[16]、贝叶斯分类(Bayes)[17]以及人工神经网络[18]等，而直观的最近邻分类是一种典型的懒惰学习算法。以下分别对这几种分类算法进行简单介绍。

1.1.3.1 决策树算法

决策树算法是以训练样本为基础的归纳学习算法，其目的是建立一种与流程图相似的树形结构。决策树的内部结点通常是属性或属性的集合，叶结点代表样本所属的类别。当分类未知类别的待分类样本时，从决策树的根结点开始对其相应的属性值进行测试，根据测试结果选择由该结点引出的分支直到决策树的叶结点。利用数据挖掘工具 WEKA 软件

(该软件可以从 http://www.cs.waikato.ac.nz/ml/weka 下载)中的 J48 算法[6],在第 1.1.2 节中所介绍的鸢尾花 Iris 数据集上建立的决策树分类模型如图 1-4 所示。

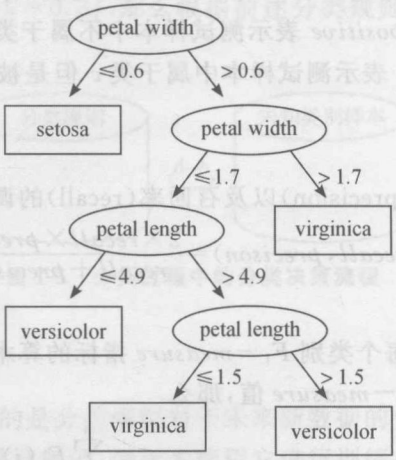


图 1-4 为 Iris 数据建立的一种决策树

第一种决策树算法是 Quinlan 在 1986 年提出的 ID3 算法[19],其使用信息增益(Information Gain)[20]作为选择分裂结点属性的标准。算法的核心步骤是:

Step1:遍历所有属性,选择信息增益最大的属性作为决策树结点,由该属性的不同取值产生该结点引出的分支。

Step2:对各分支的子集递归调用 Step1,直到所有子集只包含同一类别的样本为止。

决策树算法产生的分类规则易于理解,可解释性强。但是,算法只对较小的数据集有效,并且对噪声样本比较敏感。

1.1.3.2 支持向量机(SVM)算法

SVM 算法的基本原理是:训练样本在输入空间中可能是线性不可分的,然而可以通过一个事先选择的非线性映射 $\phi(\cdot)$,将输入样本映射到一个线性可分的高维空间。在这个高维空间中,算法基于结构风险最小化原则构造最优分类超平面 $f(x)=w^T\phi(x)+b$(w 和 b 分别是该超平面的权值向量和阈值),使得距离超平面最近的两类样本点间隔最大。所谓的"支持向量"就是那些处在超平面间隔区边缘的训练样本,这些样本对不同的类别具有良好的区分能力。虽然求解非线性映射 $\phi(\cdot)$ 的计算复杂度较高,但是由于在原空间和变换后的高维空间中只需用到向量的内积运算,因此 SVM 算法使用了核函数 $K(x,y)=\phi(x)\cdot\phi(y)$ 来简化计算。

虽然 SVM 算法在训练样本数量较少时仍然具有较好的泛化能力,并能够在一定程度