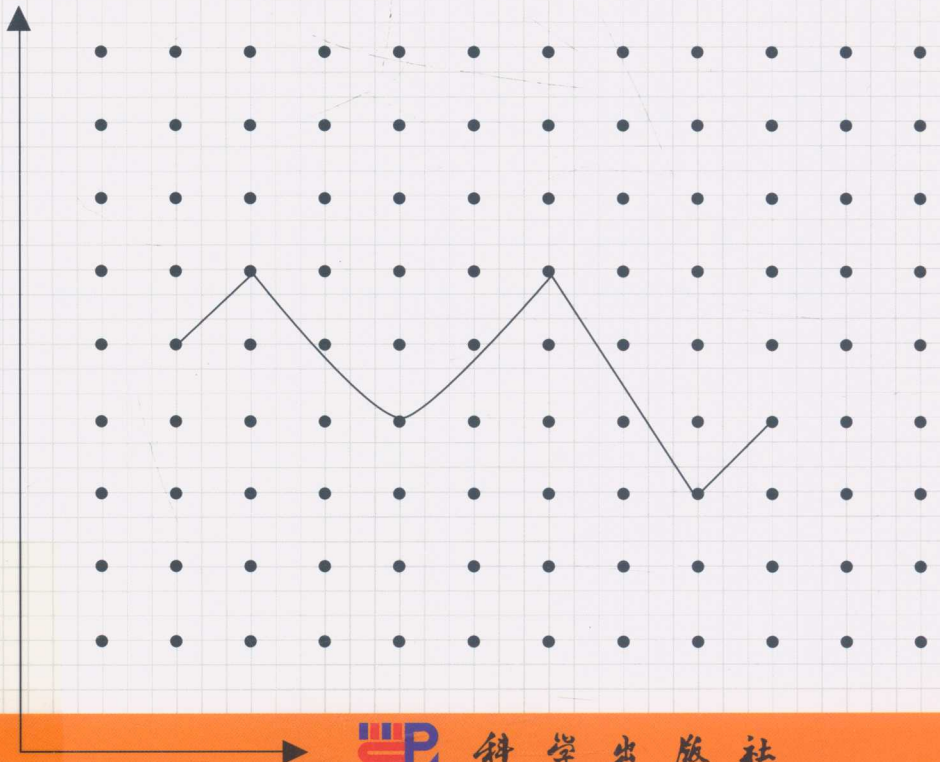
GAOWEISHUJU
DE TEZHENG XUANZE

高维数据的特征选择

——理论与算法

刘　波　何希平·著



科学出版社

重庆市检测控制集成系统工程实验室资助
电子商务及供应链系统重庆市重点实验室(重庆工商大学)资助

高维数据的特征选择
——理论与算法

刘 波 何希平 著

科学出版社
北京

内 容 简 介

特征选择是机器学习的重要研究内容，有着广泛的应用价值。特征选择主要从数据（尤其是高维数据）中选取有效特征来表示数据，从而提高机器学习算法的性能。本书以重庆工商大学等单位的机器学习、图像处理课题为基础，系统地介绍特征选择的基本概念，以及相关的理论和算法，也对它的前沿研究（如无监督特征选择）和其在计算机视觉中的应用进行详细介绍，最后对特征选择的发展方向进行展望。

本书理论联系实际，对教学、科研具有重要指导意义，可作为高等院校和科研机构从事机器学习的学者的参考书，亦可供从事大数据分析（如基因数据、计算机视觉）的专业技术人员参考。

图书在版编目(CIP)数据

高维数据的特征选择：理论与算法 / 刘波，何希平著.— 北京：科学出版社，2016.6

ISBN 978-7-03-049345-3

Ⅰ.①高… Ⅱ.①刘… ②何… Ⅲ.①统计数据-统计分析 Ⅳ. ①O212.1

中国版本图书馆 CIP 数据核字（2016）第 157941 号

责任编辑：张 展 孟 锐 / 责任校对：王 翔

责任印制：余少力 / 封面设计：墨创文化

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

成都创新包装印刷厂印刷

科学出版社发行 各地新华书店经销

*

2016 年 7 月第 一 版 开本：B5（720×1000）

2016 年 7 月第一次印刷 印张：10 1/4

字数： 205 千字

定价：59.00 元

序

机器学习是目前发展最迅速、最热门的技术之一。机器学习主要由信息获取、预处理、数据降维处理(比如选择特征)、训练模型、分析和处理数据等 5 部分组成。目前，很多训练机器学习模型算法的性能与数据维度有关。随着计算机的飞速发展，各个领域都会产生大数据，比如：基因表达数据、Web 文本数据、图像数据、金融数据等。大数据的特点就是数据维度高，数据量巨大。由于表示这些数据的特征存在冗余性、相关性，同时还有噪声，因此，用这些大数据去训练机器学习模型会导致计算效率低、容易产生过拟合等问题。特征选择是对高维数据进行预处理的重要方法。特征选择早在 20 世纪 60 年代就开始被研究，是机器学习领域中的经典问题。特征选择是基于某种评价标准，从原始特征中选择最优特征子集来表示数据。可将其看成是一种最简单的特征变换形式。

本书采用深入浅出的方式来对特征选择的基本概念、相关理论和最新研究进展进行详细介绍。对于一些抽象概念，采用图形来进行直观解释。在介绍特征选择理论的同时，也详细介绍特征选择的应用。

本书主要分为 4 部分：①特征选择的基本概念。这一部分主要介绍特征选择的定义，及特征选择的目的，同时还介绍特征变换和特征提取的基本概念，以及它们与特征选择的异同。②监督特征选择。这部分先对监督特征选择的最新进展进行介绍，然后介绍基于间隔的特征选择以及它的扩展形式：基于迹比的监督特征选择。③无监督特征选择。这部分内容首先介绍无监督特征选择的分类：过滤式无监督特征选择、绑定式无监督特征选择、嵌入式无监督特征选择。然后重点介绍嵌入式无监督特征选择算法的最新进展，对这些算法进行分类，并总结出它们的性质。目前的无监督特征选择算法主要会依赖样本的结构，而这种结构预先并不知道，因此可通过交替迭代来进行特征选择，即：先假定训练样本的结构，并进行特征选择，通过选择出来的特征确定样本的结构，一直迭代，直到最后收敛。这些方法将为进一步研究无监督特征选择算法得供很好的借签作用。④特征选择在计算机视觉中的应用。在图像分类等计算机视觉应用中，经常会对提取的图像特征进行编码，从而得到非常高维的特征向量（比如，Fisher 向量）。对这类特征，通常采用压缩的方式来减少特征数量，而压缩方式又分为：乘积量化(product quantization，PQ)方法、Hashing 的方法等。但最新的研究方法表明，对于这类特征，压缩方法会有问题，而采用特征选择能更好地提高特征选择质量。这

一部分还介绍如何用最简单的互信息特征选择方法来对 Fisher 向量进行特征选择。并由此得到结论：Fisher 向量的冗余性很少，但不相关性较大。

本书的第 1 章、第 2 章由何希平教授编写；第 3 章至第 7 章由重庆工商大学计算机科学与信息工程学院的刘波博士编写。特别感谢重庆工商大学计算机科学与信息工程学院计算机应用专业的周纯华同学为本书绘制示意图。

编写本书的过程也是我们学习的过程。为了力求做到概念准确、内容详实可靠，我们查阅了大量相关文献和资料。但由于时间和能力有很，书中内容难免不出差错。若有问题，读者可通过电子邮件 liubo7971@163.com 与我们联系，欢迎一起探讨，共同进步。

本书写作过程中得到如下实验室的基金资助：重庆市检测控制集成系统工程实验室；电子商务及供应链系统重庆市重点实验室(重庆工商大学)。

本书写作过程中得到如下项目资助：①重庆市教委研究项目“多核正则化机器学习理论研究”，项目号：KJ130709；②重庆工商大学研究项目“基于多核学习的高维数据分析研究”，项目号：2013-56-09；③电子商务及供应链系统重庆市重点实验室研究项目“基于迹比率的特征选择及关键技术研究”；④重庆市教委研究项目“大数据稀疏表示判别字典学习及其应用技术研究”，项目号：KJ1400612；⑤国家自然科学基金青年科学基金项目“真实环境下手指静脉识别与模板保护基本理论研究”，项目编号：61402063；⑥重庆工商大学研究生教改项目“基于二维码的研究生互动教学改革”，项目编号：2015YJG0205。

刘波

重庆工商大学计算机科学与信息工程学院

目 录

1 基本概念

统计机器学习(后面简称机器学习)是通过数据来建立模型并用该模型对数据进行预测和分析的过程。

机器学习主要分为两个步骤[1]：模型设计与使用模型。模型设计是指从应用环境采集数据，进行预处理，再用一些算法(如支持向量机等)对数据进行训练，并由此得到相应模型。使用模型是指对新采集的数据，采用第一步所得到的模型进行数据分析和处理。图 1-1 为高维数据的分类处理流程图。它主要由信息获取、预处理、数据降维(如选择特征)、训练模型、分析和处理数据等五部分组成。

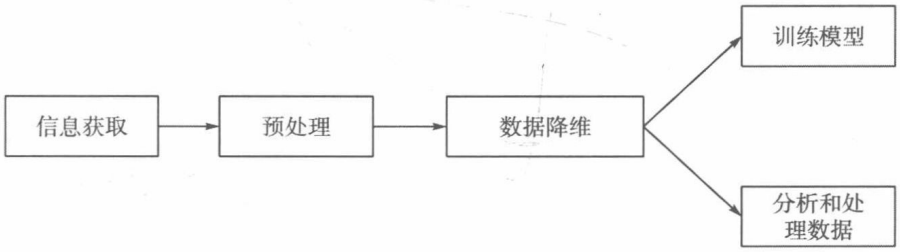


图 1-1　高维数据的分类处理流程

下面对各个步骤进行简要说明。

(1)信息获取。信息获取是指在各种应用中，利用相关的设备(如数码照相机、摄像机等)将各种对象信息转换为计算机可接受的信息并保存到存储设备中。另外，信息获取也有可能是收集用户输入的数据(如网站数据、微博数据等)。

(2)预处理。信息预处理是指将噪声从采集到的数据中去除，并对信息进行分解、合并，最后将其转化为能适合相应算法处理的数据。

(3)数据降维。在很多应用中，所采集的原始数据有很高的维数。为了让算法能高效、准确地处理这些高维数据，需通过算法对原始数据的特征进行选择。其目的是用尽量少的特征来描述原始数据，并能保持原始数据的特性。通过特征选择之后，不但可以减少算法的处理时间，还可以提高分类精度。

(4)训练模型。为了对数据进行正确处理，需要根据应用来选择合适的算法，并由此建立相应的机器学习模型。

(5)分析和处理数据。按已确定的模型对新输入数据进行判断并输出相应的分

类结果。

在上述五个阶段中，特征的好坏对后续步骤中机器学习算法的精度影响很大。

1.1 特 征 选 择

人们早在 20 世纪 60 年代就开始对特征选择进行研究[2]，它是机器学习领域中的经典问题。通常机器学习的输入是指对应用中实物或过程进行度量得到的数据，有些数据可以直接作为特征，有些则需经过预处理之后才能作为特征，本书统称这些特征为原始特征。不是所有的原始特征都有用。特征选择是基于某种评价标准从原始特征中选择最优特征子集来表示数据。设有 k 个训练样本集，则特征选择可以简单描述为：对具有 n 个原始特征的样本集 $\boldsymbol{X}$（样本与原始特征之间的关系如图 1-2 所示），给定要选择的特征数 $m(m << n)$，按某种评价标准从样本集 $\boldsymbol{X}$ 的原始特征中选择 m 个特征，从而得到新的训练样本集 $\bar{\boldsymbol{X}}$，使 $\bar{\boldsymbol{X}}$ 能最有效表达样本集 $\boldsymbol{X}$。

$$\boldsymbol{X}=\begin{array}{c} \\ f_1 \\ f_2 \\ \vdots \\ f_n \end{array}\begin{array}{c} \begin{array}{cccccc} x_1 & \cdots & x_i & \cdots & x_k \end{array} \\ \begin{bmatrix} f_{11} & \cdots & f_{1i} & \cdots & f_{1k} \\ f_{21} & \cdots & f_{2i} & \cdots & f_{2k} \\ \vdots & & \vdots & & \vdots \\ f_{n1} & \cdots & f_{ni} & \cdots & f_{nk} \end{bmatrix}_{n\times k} \end{array}$$

$$\bar{\boldsymbol{X}}=\begin{array}{c} \\ f_1 \\ f_2 \\ \vdots \\ f_m \end{array}\begin{array}{c} \begin{array}{cccccc} x_1 & \cdots & x_i & \cdots & x_k \end{array} \\ \begin{bmatrix} f_{11} & \cdots & f_{1i} & \cdots & f_{1k} \\ f_{21} & \cdots & f_{2i} & \cdots & f_{2k} \\ \vdots & & \vdots & & \vdots \\ f_{m1} & \cdots & f_{mi} & \cdots & f_{mk} \end{bmatrix}_{m\times k} \end{array}$$

图 1-2 训练数据集 $\boldsymbol{X}$ 与 $\bar{\boldsymbol{X}}$

对数据进行特征选择可降低数据的维数，特征对机器学习的影响主要表现在如下两个方面。

1. 特征对机器学习算法的精度影响

机器学习的对象是数据。通常将用于机器学习的数据划分成训练数据、测训数据和验证数据三大类。用于机器学习的数据通常用特征向量(feature vector)来表示，特征向量的每一维就是一个特征。特征对训练机器学习模型至关重要，因此机器学习的第一步就是根据具体应用选择用于训练的特征。

下面通过两幅图来说明特征对模型的重要性。

在图 1-3 中，有两类样本，分别用圆圈和三角形表示，它们的维数为 1，可通过一条竖线将这两类样本分开。对图 1-3 中的样本增加一维，这时分类直线会变成斜线，如图 1-4 所示。

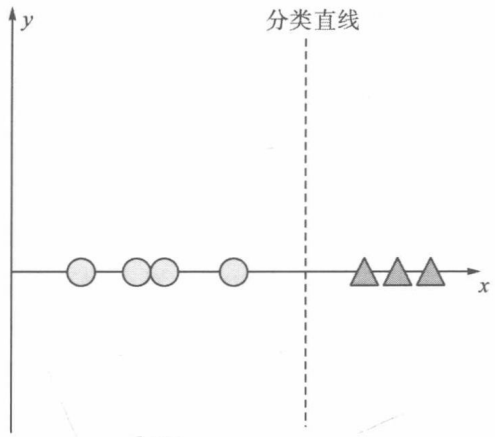


图 1-3 一维数据的分类

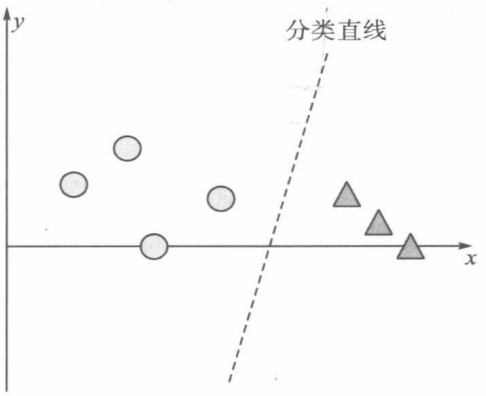


图 1-4 二维数据的分类

从图 1-3 和图 1-4 中可看出，样本的特征数量不一样，得到的分类模型很可能不一样。在图 1-4 中，若增加的一维数据与分类不相关(irrelevant)或为噪声数据，则会得到不正确的分类结果。因此，特征的数量和特征包含的有效信息对建立正确的机器学习模型至关重要。

选择好的特征不仅可提高分类精度，而且有助于得到更容易理解的算法模型。有研究结果表明，在高维数据的情况下，大部分机器学习算法所需要的训练样本数目会非常大[3-5]。而在现实中，根本无法采集到这样多的样本，从而使学习到的模型精度大大降低。Langley 等[6]的研究表明最近邻(nearest neighbor)算法的样本复杂度会随不相关特征的增加而呈指数增长。其他归纳算法也有类似的性质，例

如，对决策树而言，若特征类型为“逻辑与”，则特征数量与样本复杂度之间是线性关系，但在“异或”情况下，它们之间呈指数关系；贝叶斯分类算法虽然对不相关特征不敏感，但该算法对冗余特征很敏感[6]。

2. 提高机器学习算法的泛化能力

泛化能力(generalization ability)是指由某种学习算法得到的模型对新数据的预测能力，它是机器学习算法的重要特性[7]。在一般情况下，通过泛化误差来评价学习算法的泛化能力，也就是说泛化误差能评价一种算法得到的学习模型是否比另一种算法得到的学习模型好。一些机器学习算法的泛化误差与样本的特征数量有关，如支持向量机(support vector machine，SVM)[8]就是这样的机器学习算法。SVM 的泛化误差界可定义为[7]：以概率$1-\eta$测试样本不能被超平面间隔Δ正确分类的概率为

$$P_{\text{error}} \leqslant \frac{m}{l}+\frac{\Lambda}{2}\left(1+\sqrt{1+\frac{4m}{l\Lambda}}\right)$$

式中，l为样本数量；m为没有被超平面间隔Δ正确分类的样本数量；$\Lambda=4\dfrac{h\left(\ln\dfrac{2l}{h}+1\right)-\ln\dfrac{\eta}{4}}{l}$，这里的$h$为 VC(Vapnik-Chervonenkis)维。h与超平面间隔Δ的关系为

$$h \leqslant \min\left(\left[\frac{R^2}{\Delta^2}\right], n\right)+1$$

在这个式子中，n为样本维度，也就是特征数量，[•]表示取整数运算符。

因此，当样本维度小于$\left[\dfrac{R^2}{\Delta^2}\right]$时，特征数量越少，其 VC 维就越小，若能选择少量的好特征来表示样本，可提高 SVM 的泛化能力。

3. 特征对机器学习算法效率的影响

随着计算机的飞速发展，各个领域都会产生大数据，大数据的特点就是数据维度高，数据量巨大。下面简单介绍一些经常产生高维数据的领域。

1)基因表达数据

美国科学家于 1990 年启动人类基因计划(Human Genome Project，HGP)。该计划主要是了解生命本质、生命体生长规律、生命之间的联系、存在个体差异的原因以及认识和理解疾病产生的机制[9]。在人类遗传变异基因中最常见的是单核苷酸的多态性(single nucleotide polymorphism，SNP)，它占所有已知多态性的 90%以上。SNP 大量存在于人类基因组中，每 500～1000 个碱基对就有 1 个。据初步统计，它的数量约为 300 万个。SNP 和其他基因数据一起构成了一个大规模高维度数据。又如，蛋白质和核酸是原生质的重要成分，它们是生命的基础物质之一。

蛋白质具有催化生命体、调节生命体内的新陈代谢、抵御各类细菌入侵以及控制生命体中各种遗传信息等重要功能。在生物化学和相关的其他学科(如食品检验、临床检验等)中,蛋白质的分离和定性以及定量分析是很重要的步骤。人体蛋白质数据的维度高达 15154 维。

基因芯片又称 DNA 芯片,是研究生物基因的有效工具,它主要研究氨基酸序列(蛋白质序列)和核酸序列(DNA 和 RNA 序列)等。目前,数值型是基因芯片数据的主要表达形式,这些值以矩阵的方式存储,即基因表示矩阵。样本在不同水平下的形式是用该矩阵的一行来表示的。相同水平下所有样本的表达形式是用该矩阵的列来表示的。基因在特定条件下的表达值就是矩阵中对应的元素值。基因表示矩阵规模庞大,通常涉及数千或数万的基因数,但其样本数非常小,一般只有数十个,这是典型的高维小样本数据。基因选择是微阵列基因数据分析的核心内容,它既是建立有效分类器的关键,也是发现疾病分类的标志。目前,科研工作者正在对该问题进行探索。如何从成千上万个基因中高效地选出有效特征用于分类,一直是基因数据分析的难点[10]。

2)Web 文本数据。

随着网络的飞速发展,产生了大量 Web 网页。以新浪、搜狐等国内著名的门户网站为例,其主栏目在 100 个以上,在每个主栏目下面,又有很多的子栏目,每个子栏目下 Web 页面的内容也不尽相同。为了标注这些不同内容的网页,会产生高维数据。另外,电子邮件已成为人们相互交流和通信的一种便捷的工具,但垃圾邮件会影响人们的正常生活。据统计[11],2012 年二季度中国网民平均每周收到垃圾邮件数量为 15.3 封,中国网民平均每周收到垃圾邮件比例为 34.7%,同比上涨 1.5 个百分点。企业邮箱平均每周收到正常邮件 57.8 封,收到垃圾邮件 29.5 封,垃圾邮件占比 33.8%。普通个人邮箱垃圾邮件、垃圾信息过多影响比例为 67.7%,无法发送大附件为 66.0%,企业邮箱垃圾邮件、垃圾信息过多影响比例为 58.1%。如何区别正常邮件(ham E-mail)和垃圾邮件(spam E-mail),是一个重要的研究课题。在用词和行文格式等方面,垃圾邮件与正常邮件不一样。因此可以对邮件内容进行分析,用关键词方法基本可以有效区分垃圾邮件和正常邮件。目前市面上采用的垃圾邮件识别系统(如 Norton AntiSpam、SAproxy Pro 等)都是从每个邮件中抽取特征(也称关键词),然后采用分类算法,对这些邮件进行分类,从而识别垃圾邮件。但这些特征所构成的样本数据维度非常高,而且只有极少数特征对分类有用。

由此可见,Web 文本是高维数据。通过对这类文档数据进行特征选择,可以大大提高分类精度和处理效率。

3)图像数据

图像数据通常都是高维数据[12]。在图像数据中,人脸数据最常见。人脸识别

在公共安全、军事安全、国家安全等领域有着十分广阔的应用。同时也在智能监控、智能交通、智能门禁、公安布控中的身份识别与验证、出入境管理等领域广泛使用。人脸识别是对测试的人脸图像用训练得到的特征表示，然后进行识别。计算机识别人脸的复杂表情是一个极其困难的事情。这是因为人脸本身存在一定的弹性，会随着人的情绪而不断变化；随年龄的增长，人脸会变化(变衰老)；由于拍摄人脸的光照、成像角度和成像距离不同，所得到的人脸图像会差别很大。此外，由于图像设备的精度不断提高，在一般情况下，人脸数据可以达到几百万维，甚至上千万维像素。一般证件上人脸照片的像素也有几万，例如，如果一幅人脸图像的长和宽都是 512 像素，该图像数据为 262144 维的向量，这是非常高维的数据。人脸图像的高维性使人脸识别变得比较困难。

4)时间序列数据

股票分析、证券期货、水文气象、工业过程控制、金融、医疗诊断、科学实验等领域经常按时间顺序记录一系列数据，这些有序数据称为时间序列数据[13]。时间序列数据与静态数据不同，它是按等间距的时间段来获取数据的，其值随时间的变化而不同。对时间序列数据分析的应用十分广泛，但难度也相当大。通常如果对时间序列数据的采样频率越高或采样持续的时间越长，其数据维度越高。例如，对某个事件，用 $x_1,x_2,\cdots,x_n$ 表示在固定的时间间隔 $t_1,t_2,\cdots,t_n$ 上的取值，可用 $\boldsymbol{X}=(x_1,x_2,\cdots,x_n)$ 表示该事件。这个 n 维的向量是时间序列数据，它的维数一般很高。将经典的分类或聚类算法用于高维时间序列数据时，会大大增加这些算法的时间复杂度和空间复杂度。因此，在处理这些数据之前，需要对它们进行预处理。

5)推荐系统中用户评价数据

推荐系统[14, 15]的任务是用网站来联系用户与信息。一方面让信息能够展现在对它感兴趣的用户面前，另一方面帮助用户发现有价值的信息，引导用户获得想要的结果。最典型的推荐系统应用是电子商务领域中的 B2C(business-to-customer)模式。商家根据用户的喜好、兴趣，向用户推荐感兴趣的商品(如图书、衣服等)。在难以把握用户的需求时，如果卖家通过向用户推荐商品来满足用户的模糊需求，就可以将用户的潜在需求转化为现实需求，从而促进产品销售量。对于从事电子商务的大型网站，如 Taobao、China-pub、Amazon 等，推荐系统被大量使用。其中 Amazon 花了大约 10 年时间来研究推荐系统在电子商务中的应用。一些具有个性化服务的 Web 网站，如 IMDb 和最大的 DVD 租赁商 Netflix，也对推荐系统有很大的依赖性。推荐系统能够与用户建立长期稳定的关系，为用户提供个性化服务，对防止用户流失和提高用户忠诚度都有很大的作用。目前的推荐系统可以依赖的数据有客户活跃度信息、用户标签信息、用户对商品的反馈数据、时间上下文信息(如系统时间特性等)、社交网络等。其中用户反馈数据是用户根据对所购

商品的感受来对该商品进行评分。这是用户喜好、兴趣最真实的反映。通过该数据可以划分不同的用户群体，从而对某个用户行为的预测转换为对与该用户有相似行为的群体行为的预测。用户对商品的反馈数据由商品类别、用户爱好等构成。其中，商品类别非常多，一般有几千至几万。因此推荐系统所涉及的数据通常是高维数据。

综上所述，高维数据在各个应用领域大量存在。虽然数据高维有可能更丰富、更细致地表达事物本身，但若这些表示数据的特征有很多是噪声或不相关特征，就会给数据处理带来很许多问题。例如，高维数据中所包含的信息或结构无法被理解或展示，从而使得宝贵的“数据资源”变成“数据灾难”。另外高维数据会带来计算效率低、容易产生过拟合等问题。下面详细介绍相关特征和冗余特征的定义，然后介绍不相关特征和冗余特征对机器学习任务带来的影响。

1.1.1 相关特征

相关特征(relevant feature)通常是指对机器学习任务有贡献的特征。机器学习领域很早就在监督学习(supervised learning)中研究特征的相关性，下面给出一种相关特征的定义[3]。

定义 1.1 设第 i 个特征 f_i 的某个取值为 x_i 时，其概率 $p(f_i = x_i) > 0$，当类标签 Y 取某个值为 y 时，则 f_i 为相关特征须满足下面的条件，

$$p(Y = y \mid f_i = x_i) \neq p(Y = y)$$

从此定义可看出，若 f_i 的取值能改变对类标签 Y 的估计，则认为 f_i 是一个相关特征，否则它就是一个不相关特征(irrelevant feature)。下面可用一个例子来说明不相关特征。

在某些疾病(如癌症等)的预测中，像社保编号这样的特征就与预测任务没有关系，也就是说，社保编号不会对疾病预测产生影响，因此在这种情形下可认为社保编号是不相关特征；而年龄会对疾病预测产生影响(如年龄越大，越容易得癌症)，因此可认为年龄对疾病预测而言是相关特征。同样的特征，在不同的机器学习任务中，可能会具有不同的属性，如在与篮球有关的机器学习任务中，身高可能是很重要的特征，而在预测个人收入的机器学习任务中，身高可能就是不相关特征。

定义 1.1 给出的相关特征称为基于概率分布的强相关特征。也有弱相关特征，其定义如下。

定义 1.2 对于一组特征集 $S = (f_1, f_2, \cdots, f_n)$ 和一个单独的特征 F_i，当类标签 Y 取某个值 y 时，若 F_i 是弱相关特征当且仅当满足如下条件，即

$$p(Y = y \mid F_i, S) = p(Y = y \mid S)$$

并且存在 $S' \subset S$，使得 $p(Y = y \mid F_i, S') = p(Y = y \mid S')$。

还有其他特征相关性的定义形式(如可按复杂性度量来定义特征的相关性),有兴趣的读者可参考文献[3]。

特征相关性的检测算法通常需要与某类机器学习任务(如分类问题等)结合,本书将特征相关性检测与分类问题相结合的算法称为特征-分类检测算法。常见特征-分类检测算法有 t-检验[16]、Fisher 准则[17]、ROC[18]曲线等。为了便于解释特征相关性的基本概念,下面重点介绍两个简单的特征相关性检测算法:t-检验和 Fisher 准则,这两种算法有一定的内在联系。

t-检验也称为学生 t 检验(student t-test),它通过 t 分布理论来推论差异发生的概率,从而比较两个平均数的差异是否显著。假设样本分为两类,其类标签分别为 $y=1$ 和 $y=0$。$y=1$ 的样本数量为 n_1,相应的样本集用 X_{n_1} 来表示,相应的样本下标集为 S_{n_1};$y=0$ 的样本数量为 n_0,相应的样本集用 X_{n_0} 来表示,相应的样本下标集为 S_{n_0}。整个样本与特征之间的关系如图 1-5 所示。

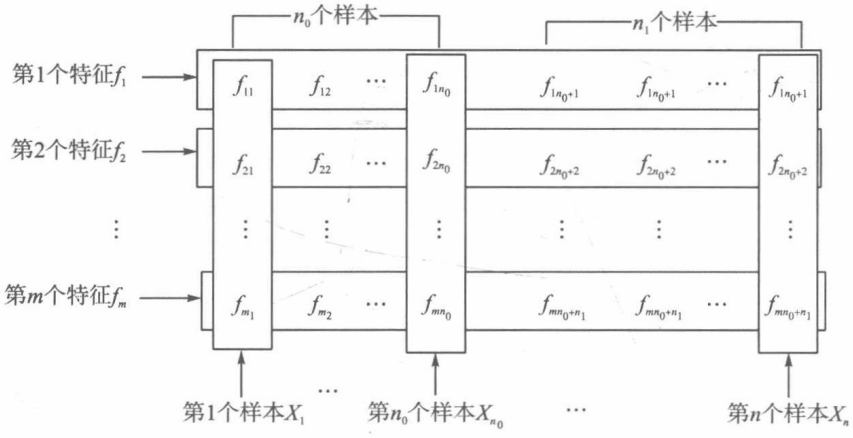


图 1-5 样本与特征选择的关系

从图 1-5 中可以看出,第 i 个特征 f_i 为第 i 行,其中 $i=1,\cdots,m$;第 j 个样本为第 j 列,其中 $j=1,\cdots,n_0+n_1$。

用 t-检验来评价这些样本中第 i 个特征 f_i 与类别的相关性是其基本思想。

(1)计算特征 f_i 在每一类上的样本均值 μ_1 和 μ_0,计算公式为

$$\mu_1=\frac{1}{n_1}\sum_{j\in S_{n_1}} f_{ij},\qquad \mu_0=\frac{1}{n_0}\sum_{j\in S_{n_0}} f_{ij} \tag{1.1}$$

(2)计算特征 f_i 在每一类上的样本方差 σ_1 和 σ_0,计算公式为

$$\sigma_1=\frac{1}{n_1}\sum_{j\in S_{n_1}} (f_{ij}-\mu_1)^2,\qquad \sigma_0=\frac{1}{n_0}\sum_{j\in S_{n_0}} (f_{ij}-\mu_0)^2 \tag{1.2}$$

(3)计算特征 f_i 的统计量 t_i,即

$$t_i = \frac{(\mu_1 - \mu_0)}{\sqrt{\frac{\sigma_1}{n_1} + \frac{\sigma_0}{n_0}}} \tag{1.3}$$

对每个特征都按上述方法计算相应的 t_i 统计量，t_i 统计量越大，则表明对应特征的相关性越强，反之则越弱。因此，可按一定标准(如设定阈值等)来去掉相关性较弱的特征，保留相关性强的特征，从而起到特征选择的作用。

可将式(1.3)中的分子部分看成两个类之间的距离，而分母的 σ_1 和 σ_0 分别反映了特征 f_i 在每一类中取值的分散程度。若希望统计量 t_i 大(这表明对应特征的相关性越强)，则需要分子尽量大，而分母尽量小，意味着要让两个类的距离尽量大，而每一类中的特征取值要尽量小。

因此，用 t-检测来评价特征相关性时，若能判断出特征的取值在类间变化较大，而在类内变化较小，则说明特征具有很好的判别性(或者说特征有较强的能力来区分不同类的样本)，即特征的相关性较强。这种思想在特征选择算法中经常用到。

但需注意，这里介绍的 t-检测只能用于训练样本的类别数为 2 的情形，多分类问题(训练样本的类别数大于 2)无法用这种方法来评价特征的相关性。对于评价多分类训练样本的特征相关性，可采用 Fisher 准则来完成。

用于判断特征相关性的 Fisher 准则的原理来自 Fisher 判别分析(Fisher discriminant analysis，FDA)方法。FDA 方法属于分类的线性模型。这种方法的基本思想是：将样本点投影到一条直线或超平面上，使得投影后同类样本的样本方差最小，而不同类样本之间的样本均值之差尽量大。前面用 t-检测来评价特征相关性时也用到了这样的思想。为了解释清楚 Fisher 准则，需先了解 FDA 方法。

对于一个多分类问题，FDA 方法会分别得到投影到向量 $\boldsymbol{w}$ 上的类间(between-class)散度矩阵 $\boldsymbol{S}_B$ 和类内(within-class)散度矩阵 $\boldsymbol{S}_W$，然后求解下面这个目标函数来得到 $\boldsymbol{w}$，即

$$\max_{\boldsymbol{w}} f(\boldsymbol{w}) = \frac{\boldsymbol{w}^{\mathrm{T}} \boldsymbol{S}_B \boldsymbol{w}}{\boldsymbol{w}^{\mathrm{T}} \boldsymbol{S}_W \boldsymbol{w}} \tag{1.4}$$

这意味着将所有样本投影到 $\boldsymbol{w}$ 以后，会使得同类样本的方差尽量小，而不同类样本之间的样本均值之差尽量大[19]。下面介绍如何得到类间散度矩阵 $\boldsymbol{S}_B$ 和类内散度矩阵 $\boldsymbol{S}_W$。

设训练样本集 $\boldsymbol{X} = [x_1, x_2, \cdots, x_n]$ 有 K 个类，第 i 个样本 x_i 投影到直线 $\boldsymbol{w}$ 上得到 y_i，即 $y_i = \boldsymbol{w}^{\mathrm{T}} x_i, i = 1, \cdots, n$。投影到 $\boldsymbol{w}$ 后的类间距离可通过每个类投影后的样本均值减去投影后的整个样本均值，然后将这些差值相加而得到，即