半监督学习

张晨光　张　燕　著

中国农业科学技术出版社

前　言

本书是国内第一本涉及半监督学习相关理论的专著，大范围而言属于机器学习范畴。机器学习是研究如何使用机器来模拟人类学习活动的一门学科，即通过机器对过往数据进行分析和处理，从而获取新知识和新技能。按照学习过程中是否有专家指导（有标记示例）机器学习可以分为监督学习、无监督学习和半监督学习。监督学习方法需要利用大量有标记的样本进行学习，而无监督学习虽然不需要有标记样本，但是缺乏先验知识的有效引导，模型的准确性难以保证。与这两种方法相比，半监督学习吸取了它们两方的优点，是介于监督学习和无监督学习之间的学习技术，它利用大量无标记示例辅助少量有标记示例进行学习，一方面充分利用了数据资源，另一方面在有标记数据稀少的情况下保证了学习效果。例如，在计算机辅助医学图像分析中，单独通过少量医学专家已经标示出病灶的图像或者其他无标示医学图像进行分析都不能得到准确的判断策略，但是结合这两种图像统一作为训练例，那就可以希望通过半监督学习技术得到比较准确的分析策略。

目前，尽管半监督学习理论和方法层出不穷，且在诸多实际应用中也取得了不菲的成绩，但它毕竟还处于发展阶段，在许多方面还存在问题。例如，基于图的半监督学习方法构图时间复杂度过高的问题；每个样本拥有多个类别标记的多标记学习问题等。针对半监督学习理论中的这些问题，本书做了以下工作：首先对以往半监督学习方法进行了总结，介绍了目前经典的几种半监督学习方法，分析了每种方法的优缺点和适用场合；提出了基于图半监督学习的图像分割方法，并针对当中构图时间过长的问题提出了哈希图半监督学习方法；针对无标记数据中某类样本数占优导致的失衡问题，提出了归一化图半监督学习方法，多个国家个人信用数据的评估实验证明了所提方法的有效性；针对多标记学习问题，在希尔伯特－施密特独立性度量方法的基础上提出了两种全新的半监督多标记学习方法，并通过图像检索、基因功能分析等多个实验证明了这几种方法的有效性；针对只有单类已标记样本的学习问题，提出了正例图半监督学习方法，将其用于图

像分割只需要单笔就能抓取所需图像部分。

本书的出版一方面可以弥补国内在半监督学习领域空白，另一方面本书提出的新方法在一定程度上也可以为半监督学习领域的发展提供一些新的思路和方法。最后，感谢海南省教育厅高等学校科学研究项目（No. Hjkj 2012-01）给予一定程度的支持；感谢北京凌云光技术有限责任公司视觉和图像系统事业部的张夏欢先生，他对本书手稿仔细审阅并提出了相关意见；感谢我的父母，他们生活上的支持让我得以安心写书。

作者

2013 年 10 月

目 录

第1章 绪 论

1.1 研究背景和意义

随着计算机科学技术特别是网络通信技术近些年的发展，人们获取信息的能力和渠道得到了极大的拓展，各行各业都积累了大量的数据。根据 Netcraft Web Server Survey 在 2012 年 8 月的统计结果，全球 Web 站点已经超过 628 170 204个，而且每天还有数以万计的新站点不断涌现。再比如，沃尔玛超市在我国拥有数百家分店，每天都为百万计的用户提供数千种商品的零售，如果将这些销售记录保存下来，那么每天需要存储的数据量可以达到几个 G 字节。海量的数据在极大丰富人们资讯的同时，也给信息的组织、查找与分析带来了极大的挑战。如何快速、准确、方便地从海量的信息库中获取感兴趣、满足需要的信息，一直是人们关心的重要课题。在各种复杂应用背景条件下，仅通过人工方式对如此庞大的数据进行分析和处理并不现实。此时，基于数据的机器学习（Machine Learning）方法就显得尤为重要。基于数据的机器学习是现代人工智能技术的一个重要研究内容和方向，其主要研究内容为：从已观测得到的样本出发，通过计算机寻找这些数据中蕴含的规律，并利用这些规律对未知数据或者无法观测的数据进行预测。目前，金融领域的信用分析、诈骗检测、股票走势预测，制造业中的控制优化、故障检测以及医疗行业的辅助诊疗和万维网上的 web 数据挖掘等都涉及机器学习理论。这一方面充分说明了机器学习的重要性，另一方面也对研究者的研究工作提出了更高的要求和挑战。

需要说明的是，虽然数据收集方法的多样化和存储技术的快速发展使得收集大量数据变得相当容易，但这些收集到的数据大多没有类别标记（性质描绘或者类别状态）的样本，而获取大量有标记的样本则相对较为困难，通常需要专家经验或者花费大量的人力物力。例如，人们可以方便地从网上获得大量图片信息，但是，只有少部分信息会有诸如天空、大海和人物等相应的类别标记，大量信息

则无任何标记。对这些无标记样本进行标记往往需要耗费大量的人力物力，甚至由于工作量巨大而完全不可行。显然，如果只使用少量的有标记样本进行训练，一方面利用它们所训练出的学习系统往往很难具有强泛化能力；另一方面，如果只使用少量“昂贵的”有标记的样本而不利用大量“廉价的”无标记的样本也是对数据资源的一种极大浪费。因此，在已标记样本较少时，如何使用大量无标记样本来改善学习性能已成为当前机器学习研究中最受关注的问题之一。在这种情况下，研究人员提出了将少量的已标记样本与大量的无标记样本一起进行学习的策略，即半监督学习方法。

目前，半监督学习已经成为机器学习与数据挖掘领域的研究热点。例如，2013 年 9 月，在谷歌学术搜索上输入“Semi-Supervised Learning ”这个关键词，搜索出来的论文就有 36 600篇。尽管如此，半监督学习作为一种新的机器学习手段，在许多方面还不完善，需要进一步研究。例如，图半监督学习需要首先构建一张图，受限于构图的时间复杂度（数据规模的立方倍），图半监督学习方法难于进入大规模数据分析领域；无标记样本中，某类的数据量远比其他类多，会使得分类面偏离正确的分类面，靠近其他数量不占优的类别，降低推广能力；再例如，一个样本同时具有多个类别的多标记学习问题，或者已标记样本中只有正类样本的单类学习问题等。

针对上面这些问题，本书首先对以往半监督学习方法进行了系统地归纳，分析了各种半监督学习方法适用的场合和缺点；其次，本书对多种传统的半监督学习方法提出了改进方案，并通过实验验证了新方法的有效性；最后，鉴于近年来文本检索，图像检索和基因功能分析领域对多标记学习方法的需求，本书在希尔伯特 - 施密特独立性度量方法的基础上提出了几种半监督多标记学习方法，并通过实验表明这几种方法可有效应用于多标记学习任务。本书的出版一方面可以弥补国内在半监督学习领域中的空白，另一方面本书提出的新方法在一定程度上也可以为半监督学习领域的发展提供一些新的思路和方法。

1.2 国内外研究现状

机器学习（Machine Learning，ML）是人工智能的一个核心研究领域，具有十分重要的研究地位，也是计算机科学中最活跃、最具研究价值的一个分支。尽管如此，对于“机器学习”至今还没有一个统一的定义，Simon 对学习定义为

“如果一个系统能够通过执行某种过程而改变它的性能，这就是学习”，而 Tom M. Mitchell 的经典著作《机器学习》[1]中将机器学习定义为“计算机利用经验改善系统自身性能的行为”。“经验”通常是指人类在过去或者现在所掌握的信息，这些信息可以是文字、图像、观测数据等，本质上都是以数据的形式存在。在“学习”过程中，计算机分析这些数据，“利用经验”获得“知识”，并对未知情况进行判断和处理。到目前为止，机器学习的研究和发展共经历 Rosenblatt 的感知器模型、学习理论基础的创立、人工神经网络以及统计学习理论四个重要的发展阶段。

基于数据的机器学习，有多种分类方法。将过往已经获得的数据组成的集合称为训练集，那么根据训练集中是否包含了“专家”的判断或者经验，机器学习的任务类型大致可以分成下面几种：无监督学习、监督学习和半监督学习。无监督学习与监督学习出现较早，发展也较为完善，半监督学习是新近提出的机器学习方法，在最近十年随着大量数据累积和计算机应用领域的不断扩展和相关理论的不断完善，也得到了很大发展。

1.2.1 无监督学习研究

无监督学习（Unsupervised learning）中没有指导者，只有输入数据，其学习的目标是发现输入数据集中潜藏的结构或者规律（这与统计学中密度函数的估计类似），因此，无监督学习通常又被形象的称为无教师归纳或无指学习。最常见的无监督学习是聚类（Clustering）问题。聚类可以描述为根据已知训练样本按照最大化簇内相似性、最小化簇之间的相似性的原则将其划分为相交或者不相交的簇的过程。通过这种方式得到的簇具有如下特性：任一样本与相同簇中的样本具有较高的相似性，而与不同簇中的样本差异较大。常见的聚类方法包括 K 均值[2]（K-means）、高斯混合模型（Gaussian Mixture Model，GMM）[3-4]等，图 1-1[5]是 K 均值聚类方法应用于图像分割的一个例子，K 是预先设定的参数值表示最终希望得到的聚类数。例如，$K=3$ 时，根据颜色值得到了三个聚类，分别是蓝色背景部分，小孩衣服以及小孩脸和黄色背景部分；K 为 2 时，实际上就是一个二值分割。除了聚类问题，无监督学习还包括异常点监测（Novelty Detection），用于发现样本点中具有明显区别的样本；维数约简（Dimensionality Reduction），保持原训练样本本质信息的同时将数据从高维空间映射到低维，常见的方法包括 PCA[6]和 KPCA[7]等。这三种无监督学习方法均没有用到已标记数据，仅

利用了无标记样本进行结构或者本维特征的学习。

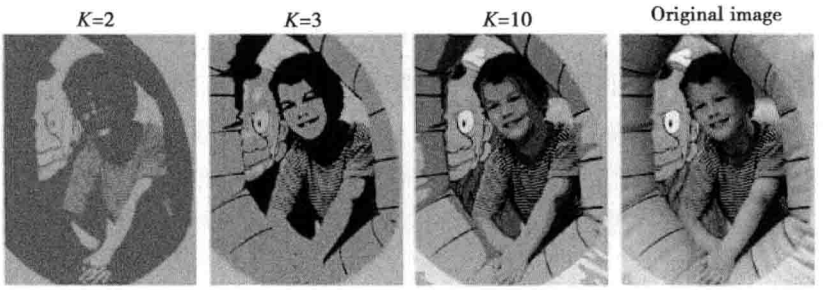

图 1-1　K-means 用于图像分割的示例，K 值表示聚类数

1.2.2　监督学习研究

监督学习（Supervised Learning）通常也称为有指导学习，常见的分类和回归问题都属于监督学习范畴。监督学习的目标是从已标记训练本学习得到样本特征到样本标记的映射关系，这种映射关系要求与已标记样本情况相吻合。映射关系和标记在分类问题中分别指分类器和类别，而在回归分析问题中就是回归函数和实值输出。需要注意的是，传统的监督学习中，通常都假设具有足够的已标记样本。如果已标记样本相对于维数或者标记数过少，那么，从中学习得到的映射会缺乏足够的泛化性，即对新样本进行判别分析的能力不足。目前的很多实际应用例如多标记学习问题，样本的维数通常都很高，且很多时候样本可能的类属也比较多，这时的已标记样本通常都相对不足，单纯依靠已标记数据显然不能得到很好的分类回归效果，一个有效的解决方法就是后面马上要提到的半监督学习方法。

常见的监督学习方法包括决策树、支持向量机、贝叶斯分类器等。决策树首次是由 Breiman 等人在 1984 年出版的 Classification and Regression Trees[8] 一书中提到，著名的 C4.5[9]、CART[10]、CHAID[11] 等算法都属于决策树方法。贝叶斯分类器通过最大化后验概率对样本类属进行判断[13]，是机器学习中的标准分类器。与基于经验风险的贝叶斯分类器不同，支持向量机建立在统计学习理论的 VC 维理论和结构风险最小原理基础上。所谓结构风险最小原理，不仅需要分类函数具有最小的经验风险，而且需要分类模型的复杂性（VC 维）尽量小[12]，从而获得更好的推广能力（Generalizatin Ability）。正因为如此，支持向量机表现出

了很多优于已有方法的性能，有望帮助解决许多原来难以解决的问题，例如，神经网络结构选择问题、局部极小点问题等。

1.2.3 半监督学习研究

半监督学习（Semi-Supervised learning）是近年来机器学习领域的研究热点，它的基本原则就是通过大量无标记数据辅助少量已标记数据进行学习，从而提高学习效果。

1.2.3.1 半监督学习方法的提出

半监督学习方法的提出一方面是因为理论上无标记样本确实有可能提高学习效果；另一方面则是因为旧有的监督学习方法只能利用已标记数据进行训练，且为了保证学习的泛化性，监督学习方法还通常需要假设已经具有足够的已标记样本。然而，实际应用中已标记数据的获取一般都比较困难，耗时、耗力且通常都需要丰富的专家经验，下面是几个例子。

（1）语音识别。样本特征就是交流中使用的语音信息，标记对应文字注解。依靠人工将语音转换成文字注解非常的费时，注解一小时的 Switchboard 电话对话就需要 400 小时。

（2）自然语言处理。样本特征是语句，标记是与之对应的语法树。Penn Chinese Treebank 中 4 000个句子的语法树的创建花费了语言专家将近两年的时间。

（3）电子邮件过滤系统。样本是电子邮件，标记指该邮件是否是垃圾邮件。垃圾邮件的内容各种各样，每天产生的垃圾邮件甚至能堵塞网络通信，通过人工方式标注的邮件内容相对而言只是当中的很小部分。

（4）视频监视。样本是视频帧，标记是视频帧的注解信息，可能是人物身份、地点名称等。众所周知，一秒视频通常就含有 24 帧，如此庞大的信息量，通过人工注解是完全不现实的。

（5）蛋白质结构预测。样本是 DNA 序列，标记是折叠结构。仅鉴别一个蛋白质结构就需花费结晶学专家几个月的时间，并且需要昂贵的实验费用。

与昂贵且稀少的已标记样本相比，无标记数据的获取相对容易，且数量众多。例如，上面提到的语音识别，通过电台就可以获得大量语音信息；再例如，自然语言处理或者视频监控需要的文本信息或者多媒体信息，在网络上随处可得。为了能够有效利用这些大量且相对便宜的无标记样本，也为了更好的利用昂贵且稀少的已标记样本，研究者们提出了半监督学习技术。在半监督学习技术提

出并受到关注的短短几年时间，研究者几乎对旧有的各种监督和无监督学习方法都做了尝试性的推广，提出了各种各样的半监督学习方法，下面简单介绍下它的主流技术。

半监督学习的思想最早可以追溯到1965年的自训练（Self-Training）方法[13]，这个方法的主要思想是利用有标号的数据进行训练得到一个分类函数，通过这个分类函数为部分没有标号的数据打上标号，再利用所有带标号的数据重复进行上述步骤直到所有的数据都得到了标号。显然，如果上述步骤中采用最小化经验风险的原则获取分类函数，那么，无标号数据对分类结果实际上并没有起到任何影响。

Dempster等人提出了半监督学习中的生成模型（Generative Model）[14]。生成模型认为每一个类的数据都服从某个混合分布，例如，高斯混合分布，理想状态下，每一类只需要一个有标号的数据就能通过其他大量的无标号数据确定它的分布，从而完成分类。相关的工作包括Nigam等人把生成模型用于文本分类，证明该方法的分类结果要比单纯的利用有标号的数据进行训练的效果要好[15]；Baluja把该方法应用于人脸的方向定位，同样具有非常好的效果[16]。生成模型也有它的缺点，它总是假设每一个类的数据服从某个分布，如果模型假设不正确，把无标号数据加入训练不但没有帮助反而会降低正确率[17]。

联合训练（Co-training）是半监督学习中的另一个方法[18]，它假设数据的特征集可以分成两个子集，且在每一子集上进行训练都能够得到一个好的分类器。两个特征子集上的分类器采取互相学习的方式完成分类工作：利用对方的分类结果重新训练，直到两个分类器对于绝大部分的数据都有一样的分类结果。相关的工作有：Jones将该方法应用于文本中的信息检索[19]；Balcan等人的试验结果表明联合训练有非常好的效果，极端情况下每个类只需要一个有标号的样本[20]。显然，联合训练的缺点在于它的假设过强。

基于图的半监督学习技术（Graph-Based Semi-supervised Learning）的关键在于它的一致性假设[21-24]：临近的点应该具有相同的标号；具有同样结构（例如，同一个聚类或流形）的点应该具有同样的标号。这两个假设，第一个反映的是局部性，而第二个反映的是全局性。直观而言，该方法是将所有的数据看成图中的点，数据之间的相似性表示点之间的边的权重，通过迭代的方式把每个点的标号信息扩展到它的近邻直至达到一个全局稳态。相关的工作包括Blum和Chawla等人的MinCut方法[23]；Zhou等人的局部全局一致性方法（The local and Global

Consistency Method)[24]等。基于图的半监督学习方法的缺点是时间复杂度比较高，计算比较缓慢。

1.2.3.2 半监督学习方法为何有用?

半监督学习方法为什么能够利用无标记数据学习？传统机器学习认为无标记数据因为没有类别标记，相当于无用的知识，因此，对监督学习的性能提高不会有任何的用处。但实际上从人类自身的认知机制而言，半监督学习是可行的。例如，一个父亲指着一条狗告知他的孩子这是一条狗，那么，这个孩子将通过大量观察到的其他样本（可能是不认识的动物，也可能是认识的动物），对什么是狗进行综合分析，进而得到是狗的准确知识。这或许不可思议，但众所周知，尽管其他的狗跟父亲告诉孩子的样子可能相差很大，但是，孩子依然很快就能准确识别狗这种动物。这方面的进一步论述，可以查看文献25，当中详细论述了人类认知领域的半监督学习机制。

下面通过图1-2从理论上简单分析下为什么无标记数据有助于学习。图中左方和右方的点分别取自两个二维正态分布，当中的正号和负号表示已经有两个样本的类属已知，分别是正类和负类，其余点的类别未知。如果不考虑无标记的样本点，那么图中实线将是监督学习方法能够得到的最好结果。然而，从图中可以看到，该分类面错误地将正类的两个样本和负类的三个样本划分到了对方那边。换个角度，按照半监督学习方法的思路考虑该问题，将得到不一样的结果。事实上，如果同时考虑已标记和无标记样本，那么虚线就是一个很自然的选择。选用虚线作为分类面，图中的所有样本都能正确分类。那么，如何解释这种现象？从统计学的角度来看，无标记数据的加入有助于样本特征的密度估计，如果可以建立样本特征与样本标记之间的联系，那么，无疑可以通过无标记样本提高学习效果。就图1-2而言，由统计学知识可知已观察到的样本的发生是合理的，即无论是已标记还是未标记样本既然发生了，那么，它们在理论上出现的概率就应该比其他情况大。为了促成这种合理性，最终的类属判断就应该使得当前观察得到的无标记和有标记数据的发生概率都尽量大。监督学习得到的实线就违背了这一原则，它没有将已经观测得到的无标记数据纳入考虑范围。按照这种划分，无法解释为什么距离负类总体如此远，而离正类总体如此近的地方会出现两个负类样本；同理也无法解释为什么距离正类体如此远，而离负类总体如此近的地方会出现正类的三个样本。

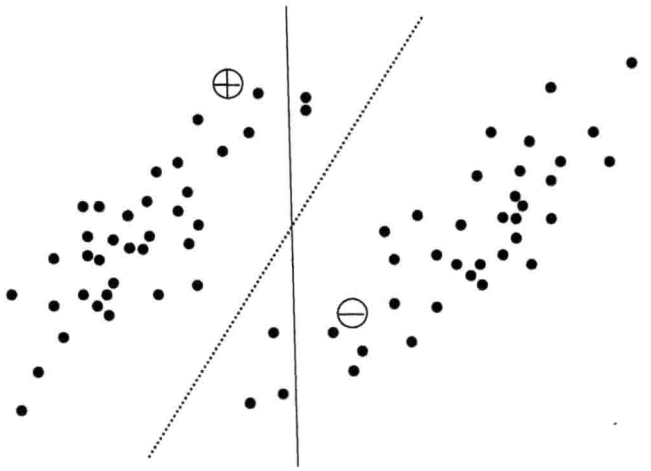

图 1-2　半监督学习示意图

注：正号表示正类样本，负号表示负类样本，其余点没有标记。所有样本取自两个二维正态分布。实线是监督学习的结果，虚线是半监督学习的结果

1.2.3.3　半监督学习方法的基本假设

一般地，半监督学习有两个基本假设，即聚类假设（Cluster Assumption）和流形假设（Manifold Assumption）。

聚类假设认为处于相同聚类的样本以较大的概率具有相同标记。聚类假设可以直观解释为如果一些对象紧凑的聚合在一起，它们就不大可能属于两种类别。聚类假设也可以表述称低密度分离假设（Low Density Separation），即决策边界应该尽量通过数据较为稀疏的区域。事实上，如果决策边界通过的是高密度区域，那么，一个聚类将会被分在了两个类别里。尽管这两种表述方式是一致的，但是因为关心侧面不一样，各自激发出了不同的方法。在聚类假设下，大量无标记样本点的作用就是帮助探明样本空间数据分布的稀疏区域和稠密区域，从而指导学习机器调整利用有标记样本学习得到的决策边界，使决策边界尽量通过数据稀疏区域。

流形假设认为高维数据总是落在低维流形上。根据流形假设，沿流形面上相近的点应该具有相似的类别标记，只要捕捉到数据所在的流形面就可以根据样本点之间的相似程度对未知样本的标记进行预测。聚类假设反映的是模型的全局特征，而流形假设主要考虑模型的局部平滑性，反映的是局部特征。在该假设下，

大量无标记样本点的作用就是让样本空间变得更加稠密，使得决策函数更好的进行数据拟合。在半监督学习中还有一个假设，称为半监督光滑假设。该假设认为，如果两个点很接近，那么他们应该具有相同的标记。基于半监督光滑假设的半监督方法为了能够利用未标记数据，通常都会通过未标记和已标记数据构建一个近邻图。在距离很小的时候，两个点的距离可以看成是他们之间的测地线距离，因此这个近邻图可以看成是对数据流形结构的一种近似，而半监督光滑假设也就可以相应看成有监督的光滑假设在流形面上的应用。

实际应用中，因为对假设的关注侧面不一样，这些半监督学习方法各有优缺点。对不用的应用场合，应该选用符合该应用环境的半监督学习方法。

1.3 研究内容和方法

本书对以往半监督学习方法进行了总结，介绍了生成模型、协同训练、图半监督学方法和半监督支持向量机，分析了每种方法的优缺点和适用场合；提出了基于图半监督学习的图像分割方法，并针对当中构图时间过长的问题提出了哈希图半监督学习方法；针对无标记数据中某类样本数占优导致的失衡问题，提出了改进图半监督学习方法，并将其用于信用评估取得了比支持向量机更好的评估效果；针对多标记学习方法，在希尔伯特－施密特独立性度量方法的基础上提出了几种全新的半监督多标记学习方法，并通过实验证明了这几种方法可有效应用于多标记学习任务；针对只有单类已标记样本的学习问题，提出了正例图半监督学习方法，将其用于图像分割只需要单笔就能抓取所需图像部分。最后的结束语总结了目前半监督学习方法发展现状，并展望了半监督学习未来的发展方向。

研究方法方面，本书采用的是理论结合实践的方法。在总结目前已有的半监督学习方法的基础上，针对其中的问题，从理论上进行改进之后通过实验对这些改进方法进行了验证；进一步区别于其他半监督学习方法的假设前提（理论基础），本书在新的假设基础上，提出了一些全新的半监督学习方法，并通过实验验证了所提方法的可行性。

参考文献

[1] Anderson J R. Machine learning: An artificial intelligence approach. 2nd edition.

San Francisco: Morgan Kaufmann, 1986

[2] MacQueen J B. Some Methods for classification and Analysis of Multivariate Observations. In Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. California: University of California Press, 1967

[3] Scudder H J. Probability of error of some adaptive pattern-recognition machines. IEEE Transactions on Information Theory, 1965, 11 (1): 363 - 371

[4] Dinov ID. Expectation Maximization and Mixture Modeling Tutorial. Statistics Online Computational Resource. http: //repositories. cdlib. org/socr/EM_ MM, 2008

[5] Tan Z, Lu R. Application of improved genetic k-means clustering algorithm in image segmentation. Education Technology and Computer Science, 2009, 2 (1): 12 - 15

[6] Jolliffe I T. Principal Component Analysis. 2nd edition. New York: Springer-Verlag, 2002

[7] Schölkopf B, Smola A, Müller K R. Kernel principal component analysis. In Artificial Neural Networks ICANN97. Berlin: Springer Berlin Heidelberg, 1977

[8] Breiman L. Classification and regression trees. 1nd edition. Boca Raton: CRC press, 1993

[9] Quinlan, Ross J. programs for machine learning. 1nd edition. San Francisco: Morgan kaufmann, 1993

[10] Burrows, William R, et al. CART decision-tree statistical analysis and prediction of summer season maximum surface ozone for the Vancouver, Montreal, and Atlantic regions of Canada. Journal of applied meteorology, 1995, 34 (8): 1 848 - 1 862

[11] Ture M, Fusun T, Kurt I. Using Kaplan - Meier analysis together with decision tree methods (C&RT, CHAID, QUEST, C4.5 and ID3) in determining recurrence-free survival of breast cancer patients. Expert Systems with Applications, 2009, 36 (2): 2 017 - 2 026

[12] Corinna C , Vapnik V. Support vector machine. Machine learning, 1995, 20 (3): 273 - 297

[13] Berger, James O. Statistical decision theory and Bayesian analysis. 2nd edition. Berlin: Springer, 1985

[14] Dempster A P, Laird N M, and Rubin D B. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B., 1977, 39 (1): 1 - 38

[15] Nigam K, Ghani R. Analyzing the effectiveness and applicability of co-training. In Proceeding of Ninth International Conference on Information and Knowledge Management. NY: ACM., 2000

[16] Baluja S. Probabilistic modeling for face orientation discrimination: Learning from labeled and unlabeled data. In Advances in Neural Information Processing Systems 11. Massachusetts: MIT Press, 1998

[17] Cozman F, Cohen I, Cirelo M. Semi-supervised learning of mixture models. In Proceedings of the 20th international conference on Machine learning. Massachusetts: AAAI Press. Semi-supervised learning of mixture models, 2003

[18] Blum A, Mitchell T. Combining labeled and unlabeled data with co-training. In Proceedings of the Workshop on Computational Learning Theory. NY: ACM, 1998

[19] Jones R. Learning to extract entities from labeled and unlabeled text. Pittsburgh: Carnegie Mellon University, 2005

[20] Chapelle O, et al. Semi-supervised learning. Massachusetts: MIT Press, 2006

[21] 张晨光, 李玉鑑. 哈希图半监督学习方法及其在图像分割中的应用. 自动化学报, 2010, 36 (11): 1 527 - 1 533

[22] 张晨光, 李玉鑑. 基于半监督学习的眉毛图像分割方法. 计算机工程与应用, 2009, 45 (21): 139 - 141

[23] Blum A, Chawla S. Learning from labeled and unlabeled data using graph mincuts. In Proceedings of the Eighteenth International Conference on Machine Learning. San Francisco: Morgan Kaufmann, 2001

[24] Zhou D, Bousquet O, et al. Learning with local and Global Consistency. In Advances in Neural Information Processing Systems 16. Massachusetts: MIT Press, 2003

[25] Zhu X, Rogers T, Qian R, and Kalish C. Humans perform semisupervisedclassification too. In Proceedings of the Twenty-Second AAAI Conference on Artificial Intelligence. Massachusetts: AAAI, 2007