于化龙 著

类别不平衡学习

理论与算法

Class Imbalance Learning

Theories and Algorithms

清华大学出版社

类别不平衡学习

理论与算法

Class Imbalance Learning

Theories and Algorithms

清华大学出版社
北京

内 容 简 介

类别不平衡学习是机器学习与数据挖掘领域的重要分支之一，其在很多应用领域中均发挥着重要作用。本书首先系统地介绍了与类别不平衡学习相关的一些基础概念及理论(第1、2章)，进而在上述理论的基础上，讨论了一些主流的类别不平衡学习技术及对应算法，具体包括样本采样技术(第3章)、代价敏感学习技术(第4章)、决策输出补偿技术(第5章)、集成学习技术(第6章)、主动学习技术(第7章)及一类分类技术(第8章)等。此外，也探讨了样本不平衡分布的危害预评估技术(第9章)。最后，对该领域未来的发展方向及应用前景做出了评述与展望(第10章)。

本书可作为高等院校与研究院所计算机、自动化及相关专业研究生的课外阅读书籍，也可供对机器学习及数据挖掘感兴趣的研究人员和工程技术人员阅读参考。

图书在版编目(CIP)数据

类别不平衡学习：理论与算法/于化龙著. —北京：清华大学出版社，2017
ISBN 978-7-302-46618-5

Ⅰ. ①类…　Ⅱ. ①于…　Ⅲ. ①机器学习　Ⅳ. ①TP181

中国版本图书馆 CIP 数据核字(2017)第 031254 号

责任编辑：许　龙　刘远星
封面设计：常雪影
责任校对：刘玉霞
责任印制：王静怡

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦 A 座　**邮　　编**：100084
社 总 机：010-62770175　**邮　　购**：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质量反馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 装 者：三河市金元印装有限公司
经　　销：全国新华书店
开　　本：170mm×230mm　**印　　张**：15.25　**字　　数**：276 千字
版　　次：2017 年 6 月第 1 版　**印　　次**：2017 年 6 月第1次印刷
印　　数：1～1000
定　　价：58.00 元

产品编号：068376-01

作者简介

于化龙(1982-)，男，哈尔滨人，博士，江苏科技大学计算机学院副教授，硕士研究生导师，东南大学自动化学院博士后。近年来，在国内外核心期刊及重要的国际学术会议上发表论文50余篇，其中被SCI与EI检索40余篇。主持包括国家自然科学基金在内的各级科研项目8项，以主要参与人身份参与国家级、省部级及市厅级科研项目多项。2015年入选江苏科技大学“深蓝学者培养工程”第二层次培养对象，2016年入选江苏省“三三三高层次人才培养工程”第三层次培养对象，是瑞士国家科学基金委科研项目特约评审专家、中国自动化学会青年工作委员会委员、江苏省计算机学会人工智能专委会委员，多个知名学术会议的PC成员，同时为20余种国内外知名学术期刊的评审人。主要研究方向为：机器学习，数据挖掘与生物信息学。

前言

FOREWORD

随着数据生成与收集技术的快速发展，如今每天在各行各业的服务器中都会新增海量的数据，这就迫使我们不得不大跨步地迈入“大数据”时代。在很多领域（尤其是商业和科研领域）的从业人员眼中，大数据犹如一座未开采的宝矿，内中裹有取之不尽的财富。而机器学习与数据挖掘技术就是那柄能开山凿路、攫取财富的“利剑”。

近年来，在产业界与学术界的双重关注下，机器学习与数据挖掘技术得到了飞速的发展，且在不断面向新应用与新挑战时，衍生出众多的新分支。类别不平衡学习便是这众多分支之一，其在机器学习与数据挖掘领域备受瞩目，很多业内主流的会议与期刊都曾以此为题举办过专刊或研讨会，如 AAAI'00，ICML'03，ACM SIGKDD Explorations Newsletter'04 以及 PAKDD'09 等。在 ICDM'05 会议上，类别不平衡问题更是被列为数据挖掘领域待解决的十大挑战性难题之一。

所谓类别不平衡问题，顾名思义，即数据集中存在某一类样本，其数量远多于或远少于其他类样本，从而导致传统的分类模型失效的问题。通常，将用于解决上述问题的算法称为类别不平衡学习算法。类别不平衡学习有着较为广阔的应用范围，如文本分类、网络入侵检测、信用卡欺诈检测、工业故障检测、软件缺陷检测、石油泄漏检测、医学诊断、药物筛选及生物信息学等。故对这一技术展开深入研究不但具有理论意义，而且还有着广泛的应用价值。

本书主要对类别不平衡学习的基本概念、基础理论及主流技术与算法展开介绍。全书共 10 章，大体上可分为以下 3 个部分：第 1 部分包括第 1，2 章，介绍类别不平衡的基本概念和基础理论；第 2 部分包括第 3～9 章，主要介绍一些用于解决类别不平衡问题的基础技术与前沿算法；第 3 部分为第 10 章，从笔者的视角对该技术未来的发展方向和应用前景做出了评述与展望。特别需要说明的是，由于此领域文献众多，初入此领域者难免会有该选读何种文献的困惑，故

笔者已将一些重要及经典的文献列出，并加以说明，置于每章后面的文献导读部分。

在此，向那些为本书出版工作提供帮助的人表达谢意。首先，感谢东南大学自动化学院的博士后合作导师孙长银教授，在东南大学做博士后的几年时间里，孙老师给了我充分的自由度，使我能安心于自己的研究课题，本书很多内容都是在这段时间研究完成的。此外，江苏科技大学的高尚教授、杨习贝副教授、王平心副教授、左欣副教授、邵长斌、郑尚、秦斌、徐丹、鞠恒荣、洪淑芳、袁玉龙、杨菊、李青雯、席晓燕，东南大学的杨万扣副教授、刘金花、姚乔兵，天津大学的穆朝絮老师以及美国爱荷华大学的倪军副教授等均在本书出版过程中给予了支持与帮助，在此一并表示感谢。

其次，感谢国家自然科学基金(No. 61305058)、江苏省自然科学基金(No. BK20130471)、国家博士后特别资助计划项目(No. 2015T80481)、国家博士后基金(No. 2013M540404)、江苏省博士后基金(No. 1401037B)、江苏省教育厅高等学校自然科学研究项目(No. 12KJB520003)及江苏科技大学“深蓝学者”计划培养基金对本课题研究工作及本书出版工作所提供的经费支持。

笔者深知自己才疏学浅，对类别不平衡学习技术仅可做到管中窥豹，且鉴于时间与精力有限，成稿仓促，书中难免会有错误与疏漏之处，望读者不吝指出，笔者将不胜感激。

笔　者

于江苏科技大学

目 录

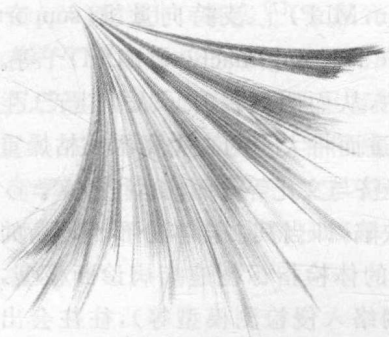
CONTENTS

第1章

绪　论

第1章

绪 论



1.1 引言

近年来，随着数据获取与数据存储设备价格的不断下降，各行各业所记录与积累的数据量也在大幅增长，其催生了一个新的名词——大数据(big data)。可以说，大数据是一把双刃剑，它在为信息产业快速发展带来机遇的同时，也向现有的信息技术提出了强有力的挑战。若信息产业的从业者应对得当，则将如徐子沛先生所言："大数据将成为人类文明新的土壤，在这片土壤之上，人类将开始建设一个智能社会"[1]。

那么，如何从纷繁复杂的大数据中挖掘出有用的知识，并用其来加速科学技术发展及辅助商业决策呢？这便需要用到两项关键技术：机器学习与数据挖掘，二者相互关联且密不可分。其中，机器学习主要为数据挖掘任务提供模型与工具，故其更加偏重于理论；而数据挖掘则更多地关注于不同数据类型的特点及不同领域和层面用户的实际需求，并根据这些需求向机器学习寻求理论帮助，换言之，数据挖掘更加注重实际应用。近几年，随着大数据概念的不断升温，机器学习与数据挖掘技术也已逐渐成为科学界及产业界关注的焦点[2]。

一直以来，分类技术(classification)都是机器学习与数据挖掘研究领域的重要组成部分，其流程为：首先，通过数据采集设备收集一定数量的样本(如图像、文本、视频、数字等)并进行简单的预处理操作；然后，由有经验的领域专家为这些样本手工添加类别标注；继而，利用上述已添加类别标注的样本训练一个分类模型，原则在于要保证该模型可最大限度地区分不同类别的样本；最后，采用已训练好的分类模型对未知样本的类属进行预测。截至目前，已出现了大量的

分类模型构建方法，较为常用的有：决策树(decision tree)[3]、K 近邻(K-nearest neighbors，KNN)[4]、朴素贝叶斯(naive Bayes)[5]、Logistic 回归(Logistic regression)[6]、多层感知器(multilayer perceptron，MLP)[7]、支持向量机(support vector machine，SVM)[8]及极限学习机(extreme learning machine，ELM)[9]等。分类技术也已被成功地应用于很多实际领域中，从而明显提升了人类生产、生活、交通、安防及医疗卫生等领域的智能化水平，进而将大量的人力资源从枯燥重复的劳动中解脱出来，提高了人类社会在政治、经济与文化等方面的运行效率。

然而，传统的分类技术通常存在一个致命缺陷，即当其在样本分布不均衡的数据上训练时(如采用 99 个健康人和 1 个病人的体检指标创建疾病诊断模型，99990 个正常数据包和 10 个病毒数据包构建网络入侵检测模型等)，往往会出现分类面偏倚的现象，从而无法得到理想的分类效果，在严重情况下，模型甚至会完全失效。上述问题在机器学习与数据挖掘领域通常被称为“类别不平衡”(class imbalance)问题，人们也习惯地将用于解决上述问题的算法统称为类别不平衡学习算法[10-12]。自 20 世纪 90 年代末以来，类别不平衡学习一直是机器学习与数据挖掘领域的研究热点与难点之一，很多业内的主流会议与期刊也都曾以此为题举办过专刊或研讨会，如 AAAI'00[13]，ICML'03[14]，ACM SIGKDD Explorations Newsletter'04[15]以及 PAKDD'09[16]等。在 ICDM'05 会议上，类别不平衡问题更是被列为数据挖掘领域待解决的十大难题之一[17]。时至今日，学术界及产业界对该问题的研究热情仍未消退，且随着大数据的出现而呈现逐渐升温的态势。图 1-1 为采用“class imbalance ∩ classification”作为关键词在工程索引(EI Village)中查询到的近年来该领域发表文献情况。

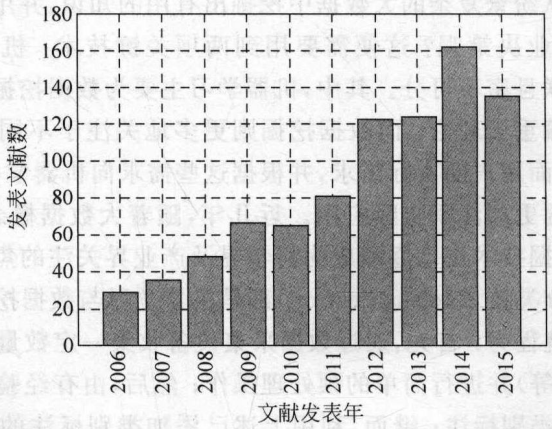


图 1-1 在 EI Village 中以“class imbalance∩classification”作为关键词所检索到的近年文献发表情况

从图 1-1 中可以看出，尽管存在轻微的波动，但在近 10 年，该领域的文献发表数仍表现出逐年递增的趋势，特别是在 2012 年后，每年发表的文献数均保持在 120 篇以上。事实上，考虑到关键词及文献库选取的局限性，上述统计结果是在打了严重折扣的情况下得到的，实际上要远多于此。由此可见，类别不平衡学习已逐渐发展成为机器学习与数据挖掘领域的重要分支之一。

在本章的后续部分，将陆续对类别不平衡问题的基本概念、常用的类别不平衡学习技术及类别不平衡学习所适用的应用领域做概要性的介绍，以使读者能对类别不平衡学习有初步简单的了解与认识，为后续章节的学习打下坚实的基础。

1.2 基本概念

类别不平衡就是指在分类任务中不同类别的训练样本数目差别很大的情况。不失一般性，我们可以假设训练集中只包含两类样本，即待处理的分类问题为二分类问题，同时，为了保证更好的可视化效果，不妨设每个样本均具有两个特征。图 1-2 给出了平衡样本集及不平衡样本集的对比效果，其中：平衡样本集中两类各有 500 个样本，类别 1 样本在特征 1 的[0,0.7]取值区间及特征 2 的[0,1]取值区间上分别服从均匀分布，而类别 2 样本在特征 1 的[0.5,1]取值区间及特征 2 的[0,1]取值区间内服从均匀分布；不平衡样本集同样包含 1000 个样本，但类别 1 被分配 900 个样本，而类别 2 仅有 100 个样本，其各自的分布与平衡样本集完全一致。

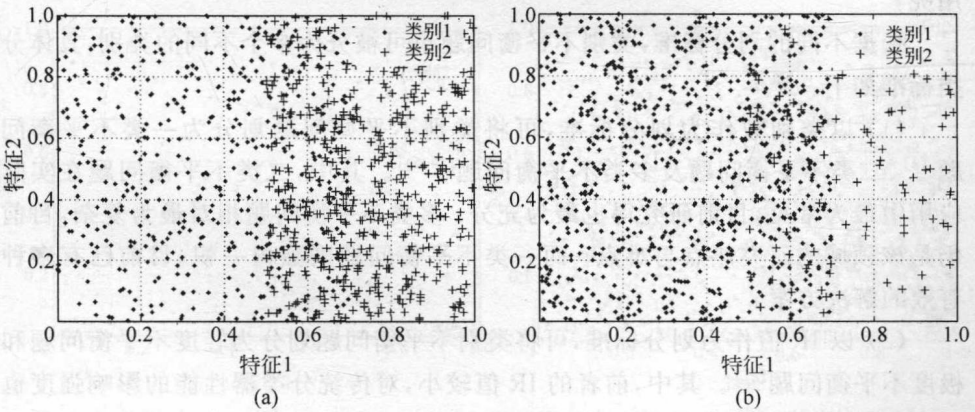


图 1-2 平衡样本集与不平衡样本集的对比效果

(a) 平衡样本集分布；(b) 不平衡样本集分布

从图 1-2 中不难观察到一个有趣的现象，即仅从视觉效果来看，在平衡与不平衡训练集上，两类样本具有完全不同的分割位置，即表明它们的分类边界不同。但据我们所知，在这两个训练集上，同类样本的分布是完全一致的。那么，这难道仅仅是由于人类视力的缺陷而引起的错觉么？事实却非如此，因为绝大多数传统的分类算法都犯了和人类眼睛同样的错误。

众所周知，尽管传统的分类算法在构造机理上各不相同，但却几乎均遵循一个共有的原则，即训练误差最小化原则。在平衡训练集上，采用训练误差最小化原则无疑会得出最优结果，而当训练集为不平衡时，若仍然坚持采用这一原则则会产生严重后果。回顾图 1-2(b)，可清晰地观察到两类样本在特征 1 的[0.5，0.7]取值区间内相互交叠，且在这一区间内类别 1(多数类)的样本个数远多于类别 2(少数类)，若采用训练误差最小化原则，则处于此区间内的少数类样本均会被误判，导致少数类的分类精度远低于多数类，从而致使所训练分类模型的质量大打折扣，甚至完全失效。这便是类别不平衡问题对传统分类算法所提出的挑战。

在类别不平衡问题中，人们习惯将包含样本数较多的类别称为负类(negative class)，而将样本数较少的类别称为正类(positive class)。此外，另一个较重要的概念为不平衡比率(imbalanced ratio，IR)，它的值为负类样本数与正类样本数之比。通常，IR 值越大，其对传统分类器性能的危害也会越大。考虑一个 IR 值为 99 的训练样本集，若在构造分类器时，将所有的正类样本均误判为负类，其分类精度仍可达到 99%，而这样的精度对于建立在训练误差最小化原则之上的传统分类算法而言，是绝对可以接受的，但这样的分类模型又有什么用呢？

根据不同的划分标准，类别不平衡问题也可被分为多个不同的类别，具体分类标准如下：

(1) 以类别数作为划分标准，可将类别不平衡问题划分为一类不平衡问题[18]、二类不平衡问题及多类不平衡问题[19,20]。其中，二类不平衡问题在实际应用中最为常见，目前研究得也最为充分；多类不平衡问题相对最为复杂，目前仍是该领域的研究热点与难点；而一类不平衡问题则独树一帜，目前已有多种有效的解决方案。

(2) 以 IR 值作为划分标准，可将类别不平衡问题划分为轻度不平衡问题和极度不平衡问题[21]。其中，前者的 IR 值较小，对传统分类器性能的影响强度也不大，而后者则会对传统分类算法构成较大威胁，极端情况下会令其完全失效。

(3) 以作用范围作为划分标准，可将类别不平衡问题划分为类内不平衡问题与类间不平衡问题。其中，前者又被称作类内子聚集或小析取项问题，其主要