BUJUNHENG SHUJU
SVM FENLEI SUANFA
JIQI YINGYONG

不均衡数据SVM分类算法及其应用

陶新民 刘福荣 杜宝祥 著

黑龙江科学技术出版社

不均衡数据SVM分类算法及其应用

陶新民 刘福荣 杜宝祥 著

黑龙江科学技术出版社

图书在版编目（CIP）数据

不均衡数据SVM分类算法及其应用 / 陶新民，刘福荣，杜宝祥著. --
哈尔滨 ：黑龙江科学技术出版社，2011.10
ISBN 978-7-5388-6832-6

Ⅰ. ①不… Ⅱ. ①陶…②刘…③杜… Ⅲ. ①向量计算机－算法理论
Ⅳ. ①TP301.6

中国版本图书馆CIP数据核字(2011)第223685号

责任编辑 王姝

不均衡数据SVM分类算法及其应用

BUJUNHENG SHUJV SVM FENLEI SUANFA JIQI YINGYONG

陶新民 刘福荣 杜宝祥 著

出　　版　黑龙江科学技术出版社
　　地址：哈尔滨市南岗区建设街41号 邮编：150001
　　电话 0451－53642106　传真 0451－53642143（发行部）
发　　行　全国新华书店
印　　刷　黑龙江省地质测绘印刷中心印刷厂
开　　本　850mm×1168mm　1/32
印　　张　11.125
字　　数　240 000
版　　次　2011年10月第1版·2011年10月第1次印刷
书　　号　ISBN 978-7－5388－6832-6/TP·121
定　　价　35.00元

编者的话

支持向量机算法（SVM）是以统计学习理论为基础的一种机器学习方法，它以其扎实的理论基础以及完整的理论推导，成为处理小样本学习、非线性、局部极小值等问题的有效工具。支持向量机与神经网络相比，具有收敛速度快、稳定性好和泛化能力强等优点，目前广泛应用在机器学习的各个领域。然而在现实应用中，由于异常样本不容易获得，因此导致可利用的训练样本数目往往是不均衡的。虽然传统的支持向量机算法具有处理小样本数据集的优点，但是在面对不均衡训练数据时，分类效果很不理想，为此很多学者对支持向量机算法进行了改进。改进方向主要归纳为以下两类：一是从算法角度，即改进分类算法，主要是通过改变概率密度、调节各类样本之间的代价函数、单类识别方法、组合调整分类边界等措施使其更有利于少数类的分类；另一个是从数据集的角度，即重构数据集，其中包括过抽样和欠抽样，该方法是通过数据预处理手段来解决不均衡数据问题。

本专著首先分析了 SVM 分类算法的机理，并阐述了传统 SVM 算法处理不均衡数据集时分类效果不理想的原因，介绍了目前最流行的基于数据预处理的不均衡数据分类方法：其中包括自适应过抽样算法、边界人工样本生成算法、基于聚类的欠抽样算法以及基于集成的智能欠抽样算法，同时还包括作者课题组提出的基于谱聚类的欠抽样算法和基于阴性免疫的过抽样算法；另外我们还详细介绍了集成分类算法，其中包括对集成算法的理论分析以及两类不均衡数据代价敏感学习算法的阐述；接着作者结合自身的工作介绍了一种基于谱聚类欠抽样的

SVM 不均衡分类算法、基于核聚类集成的 SVM 不均衡分类算法以及基于主动学习欠抽样的不均衡数据 SVM 分类方法，最后给出了不均衡数据分类领域未来的发展方向及其面临的挑战。本书的内容不仅包括作者个人的工作，同时涵盖了作者及其历届研究生共同研究的成果。

特别应当指出的是，本书的编写得到了国家自然科学基金面上项目(61074076)、教育部博士学科点专项科研基金(20092304120017)、中国博士后科学基金面上项目(20090450119)以及黑龙江省博士后基金项目(LBH-Z08227)的资助，同时还得到了黑龙江科学技术出版社专家和编辑的鼓励。哈尔滨工程大学的付丹丹同学详细阅读了全书，并提出许多宝贵的意见，黑龙江科技学院的徐晶老师对原稿中的一些不足和欠妥之处给予了补充和纠正。郝思媛和张冬雪同学则负责书中部分绘图和翻译工作，并完成了全部书稿的整理工作，在此向他们表示衷心的感谢。

同时，本书之所以能够得以完成，与爱妻薛子恒女士和爱女陶思睿在精神上的长期支持与鼓励是分不开的，在此深表谢意。

在此书的编写过程中，作者参阅了国内外许多有关不均衡数据分类方面的资料，从中吸取了新的思想，新的内容，同时又力图有所突破，有所创新，然而不均衡数据分类是近年来发展起来的新兴方向，可供参阅的资料不多，加之作者能力和水平有限，时间仓促，书中难免会有很多错误和不足之处，敬请阅读本书的老师和同学予以指正。

陶新民

2011 年于哈尔滨

目 录

第一章 概 述

随着网络技术的飞速发展，计算机已经具备了存储和处理海量数据和访问远程站点的能力。以一家连锁超市为例，若它在全国具有数百家分店，那么它可以通过计算机网络为数百万客户提供商品零售服务。其中每个销售点都记录并存储每笔交易数据：交易的日期、购买的商品、消费总额等等，并通过网络进行汇总。接下来商家会对这些数据进行分析，并将它转化为可用信息，如通过分析客户购买商品时的关联为下一阶段商品位置的摆放提供依据等等。这个例子可以说明在实际生活和工业生产中，观测样本的获取并不是一个完全随机的过程，而是存在某种潜在的规律，比如顾客在超市购买商品并不是完全随机的：买啤酒时一般也会购买薯片；夏天会买冰激凌；冬天会买热水袋等等。这些统计样本中往往存在着某种确定模式[1]，然而如何从这些数据中挖掘出有用信息，并为下一步的规划提供依据则是亟待解决的首要问题。

机器学习的目的就是将已知的观测样本转换成模型，解决样本信息的有效利用问题。其主要研究方法是从观测样本出发寻找规律，运用这些规律对无法观测或者未知的样本进行预测，其最终目标是使机器具有良好的泛化能力[2]。机器学习也是人工智能的组成部分，它还可以解决视觉、机器人以及语音识别等方面的很多问题。随着各行各业大量样本的涌现，如何利用这些样本提高管理水平、增加企业效益、保障社会与信息安全，成为当前社会不得不解决的重要问题，所以机器学习的发展得到了社会和学术界的高度重视[3]。值得一提的是，在2005年的国际人工智能会议(IJCAI’05)收录的文章中有近一半是与机器

学习的研究相关的[4]。

但是，我们面对的观测样本与传统意义的样本并不相同。传统的样本一般是从精心设计的实验装置中筛选出来的，这些样本往往满足一定条件；然而，我们所获得的观测样本往往具有涌现性，比较典型的如网络样本、金融样本以及生物样本，对于这些样本，我们无法有效控制其产生过程[5]。这就是说虽然我们得到大量的样本，但其中对我们有用的样本往往很有限，通常仅占全部样本的一小部分。这种某类样本的数量明显少于其他类样本数量的样本集称为不均衡样本集。在以往机器学习的研究中，通常都假设用于训练的样本集是均衡的(即各类所含样本数目大致是相等的)，这就导致传统机器学习方法在面对不均衡数据集分类任务时往往泛化性能不高[6]。一方面不均衡样本问题大量存在于人们的现实生活和工业生产之中，比如网络入侵、信用卡欺诈检测、医疗检测、语音信号处理、文本分类和信息检索等[7]；另一方面不均衡样本给传统机器学习的研究带来了新的挑战，所以对不均衡样本的研究成为了学者们争相解决的问题。由于不均衡样本问题中各类学习样本数量不均衡，这就使得以精度为评价标准的学习方法不能很好地处理此类问题。现以二分类为例，假若两类学习样本的数量比为1：99，那么通常的模型都会倾向于将所有的样本归属为数目多的一类(多数类) [8]。由此可得到高达99%的分类精度。然而这样的结果在实际应用中并不能完全令人满意，尤其当少数类样本信息具有很高利用价值时更是如此，所以对不均衡样本下的机器学习的研究成了最近几年热门的课题[9,10]。

近期一些重要的学术会议也对不均衡样本分类进行讨论与分析，例如：由美国人工智能协会主办的关于学习不均衡样本

集的研习会(AAAI’00)[11]，不均衡样本集机器学习的国际会议研讨会(ICML’03)[12]，计算机机械专家组在知识发现和样本挖掘探索活动的协会(ACMSIGKDD 探究’04)[13]。对不均衡样本学习问题的关注和研讨活动的进行促进了该研究领域的快速发展。图1.1估计了自20世纪90年代以来有关不均衡学习问题的论文发表情况，该数字已超过了十年前基于电气和电子工程学会(IEEE)和计算机协会(ACM)样本库中的统计量。可以看出，这个领域上的有关论文呈现明显的增长趋势。由于该领域相对比较新且发展迅速，因此将该领域近年来的研究成果进行整理和综述就显得十分必要，这也正是本书出版的目的。由于作者近几年一直从事不均衡数据分类算法的研究并提出了多个有效算法，因此作者希望出此专著与学者们共同切磋。该书不仅介绍了同行及作者自己的研究工作，而且对下一步的研究也进行了展望，以供学者们共同研究探讨。

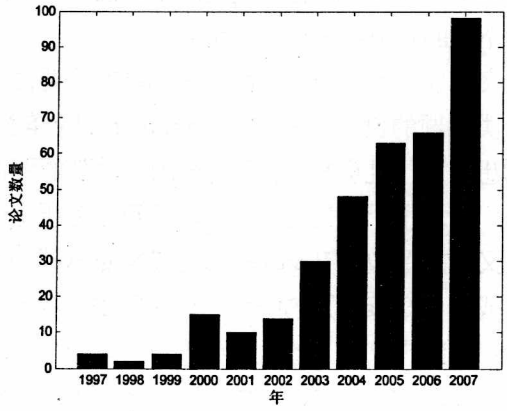

图 1.1 不均衡样本学习问题方向的研究论文发展趋势

1.1 问题的本质

从技术角度上说，任何在不同类之间展现出不等分布的样本集都应被认为是不均衡的。但是，业界普遍认为不均衡样本对应的样本集应该展现出明显的不均衡特征，特殊情况下，甚至需要表现出极度不均衡的特征。具体来说，这种不均衡形式被称为类间不均衡；常见的多数类与少数类样本比例是100∶1，1000∶1，10000∶1，在上述情况下，多数类比少数类的样本个数要多很多[14-16]。需要说明的是，这样的描述并不意味着类间不均衡问题只存在于二分类问题中，其实在多分类问题中也同样存在。当然本书主要集中讨论两类不均衡学习问题，对于多类别的不均衡学习问题也做了简单介绍，但是由于多分类问题可以转换为多个两分类问题，因此这里不做过多讨论。另外需要注意的是，在多数类与少数类出现类间不均衡的同时，在多数类别样本中也可能同样存在类内的不均衡[17-22]，两个概念我们将在接下来的几节中做详细介绍。

为了突出现实世界中不均衡学习问题的重要性，我们列举一个二分类问题的例子。考虑“Mammography样本集”，就是对不同的患者进行乳房X线检测获得的一系列的图像，这个样本集已经广泛地用于不均衡学习问题的算法分析上[23-25]。分析二进制意义下的图像，“Positive”或是“Negative”分别表示“癌症病人”或是“健康人”的图像。在实验中，我们希望未患癌症人的数目要比患癌症人的数目要大得多；事实上，样本集包含10923个“Negative”(多类)样本和260个“Positive”(少数类)样本。理想情况下，我们训练的目的是为了使得到的分类器在少数类和多数类上都能提供100%的预测精度。事实上，我们发

现分类器趋向于给出一个严重不平衡的精度，其中多数类有接近100%的精度，而少数类的精度却在0—10%的范围内，例如文献[23-25]。假设分类器在mammography样本集少数类的精度是10%，分析可知，这就暗示了有234个少数类的样本被错分为多数类样本。这个结果等价于有234个癌症患者被误诊为健康人。在医学界，这种错分情况的后果将是非常严重的，要比将健康人诊断为癌症患者的损失要大得多[26]。显而易见，我们需要的分类器应该是在不严重损坏多数类精度的情况下，在少数类上获得尽可能高的精度。此外，这也暗示了使用单一的评价准侧，例如全局精度或是误差率，是不能给不均衡问题提供足够的评价信息的。因此利用含有更多信息的评价指标，例如接收机特性曲线、精度—recall曲线和代价曲线，对不均衡样本分类性能进行评价是非常必要的，在本章我们将详细讨论这些问题。上面阐述的仅是不均衡分类问题在医学领域的应用，通过进一步推断，我们会发现这种情况也同样适用于欺诈检测、网络入侵和石油泄漏检测等分类领域[25-29]。

上面描述的不均衡形式常常被认为是内在的，即样本的不均衡是由样本空间的性质导致的。研究发现，不均衡样本不仅仅限制于内在的变化，可变因素如时间和存储量，也可能导致样本集不平衡，这种类型的不均衡被认为是外在的，即不均衡与样本空间的性质不是直接相关的。外在不均衡与内在不均衡一样值得关注，这是因为即使样本空间本身是均衡的，也有可能最终因某种原因导致收集到的样本集不均衡。例如，假设连续的平衡样本流在特定的时间间隔上产生一个样本集，如果在这个间隔内，传输中有零星的中断就会导致样本不能传输，那么最终所收集的样本集很有可能是不平衡的，在这种情况下，