



Heilongjiang Science & Technology Press

BUJUNHENG SHUJU

SVM FENLEI SUANFA

JIQI YINGYONG

不平衡数据

SVM 分类算法及其应用

陶新民 刘福荣 杜宝祥 著



YZLI0890146421

黑龙江科学技术出版社

不平衡数据 SVM 分类算法 及其应用

陶新民 刘福荣 杜宝祥 著



YZLI0890145421

黑龙江科学技术出版社

图书在版编目 (C I P) 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 SHUJY 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, 目前广泛应用于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	1
1.1 问题的本质	4
1.2 国内外不均衡学习研究现状	11
1.2.1 算法层面的处理方法	12
1.2.2 样本层面的处理方法	13
1.3 评估指标	14
1.4 本书的安排	17
第二章 支持向量机综述	19
2.1 支持向量机	20
2.1.1 最优分类界面的定义	20
2.1.2 最优分类界面的构建	23
2.1.3 广义最优分类界面	26
2.1.4 支持向量机的构建	29
2.2 核函数	30
2.2.1 高斯核函数	32
2.2.2 多项式核函数	32
2.2.3 S型核函数	32
2.3 不均衡数据对 SVM 性能的影响	33
2.4 本章小结	35
第三章 不均衡学习的抽样方法	37
3.1 随机过抽样和欠抽样	37
3.2 informed 欠抽样	38

3.3 数据生成的合成抽样方法	43
3.4 自适应合成抽样方法	45
3.5 利用数据清洁技术的抽样	50
3.6 基于聚类的抽样方法	53
3.7 抽样和 Boosting 算法的集成	56
3.8 实验分析	58
3.9 本章小结	68
第四章 基于 ODR 和 BSMOTE 的不均衡 SVM 分类算法	71
4.1 KNN 算法	71
4.2 ODR 欠抽样算法	73
4.3 ODR-BSMOTE-SVM 算法	75
4.4 对比算法简介	77
4.4.1 RU-BSMOTE-SVM 算法	77
4.4.2 KSMOTE-SVM 算法	79
4.5 仿真实验及性能分析	80
4.5.1 实验数据	80
4.5.2 不同算法的性能比较	81
4.5.3 不同比例下不均衡数据的性能比较	84
4.5.4 参数 α 对算法性能的影响	86
4.5.5 分析结论	90
4.6 本章小结	90
第五章 基于阴性免疫过抽样的不均衡分类算法	93
5.1 不均衡数据下传统过抽样算法	94
5.2 基于阴性免疫的过抽样算法	95
5.2.1 检测器表示及亲和度定义	95
5.2.2 检测器移动操作	96
5.2.3 检测器间覆盖度的测量标准	96

5.2.4 检测器克隆增殖操作	97
5.2.5 检测器变异操作	97
5.2.6 检测器克隆选择及消亡操作	98
5.3 基于阴性免疫的过抽样算法流程	99
5.4 对标称值属性的处理	101
5.5 仿真实验及性能分析	101
5.5.1 NI 算法生成人工样本的试验	101
5.5.2 不同过抽样算法的性能对比试验	103
5.5.3 不同不均衡数据比例的性能对比试验	105
5.5.4 分析与总结	106
5.6 本章小结	108
第六章 基于谱聚类欠抽样不均衡 SVM 分类算法	109
6.1 SVM 在不均衡数据下分类边界的偏移	110
6.2 基于谱聚类的欠抽样算法	112
6.2.1 传统欠抽样算法分析	112
6.2.2 基于谱聚类的欠抽样	113
6.3 基于谱聚类欠抽样不均衡 SVM 算法	115
6.4 实验分析及对比	117
6.4.1 实验数据	117
6.4.2 不同算法的分类性能比较	117
6.4.3 不同比例下不均衡数据分类性能比较	129
6.4.4 高斯核半径的参数对算法性能的影响	129
6.5 本章小结	134
第七章 集成方法	137
7.1 分类器集成学习	138
7.2 Bagging	141
7.3 随机森林	143

7.4 Boosting.....	143
7.4.1 AdaBoost 算法.....	145
7.4.2 上边界定理及参数 α 的选择	148
7.4.3 加权的效率.....	151
7.4.4 前向的 stagewise 累计模型.....	152
7.5 本章小结	153
第八章 集成算法的理论分析	155
8.1 平方损失函数的混淆分解	155
8.2 偏差和方差的错误分解框架	156
8.3 错误分解与混淆分解的关联	157
8.4 多数投票策略集成的错误分解	160
8.5 差异性创建策略	162
8.5.1 训练样本的处理	162
8.5.2 结构的处理	165
8.5.3 惩罚项方法	165
8.5.4 进化方法	168
8.6 不同训练集下分类器的偏差和方差错误测试方法.....	171
8.7 SVM 集成算法性能随核参数变化试验	174
8.8 不同分类器之间的差异性度量方法	178
8.9 采样率对分类精度的影响试验	180
8.10 本章小结	185
第九章 两类不均衡数据学习的代价敏感学习算法	187
9.1 代价敏感的 AdaBoost 算法	189
9.1.1 AdaC1 算法.....	191
9.1.2 AdaC2 算法.....	193
9.1.3 AdaC3 算法.....	196
9.1.4 三种算法性能分析	199

9.2 成本敏感指数损失和 AdaC2	201
9.3 代价因子分析	202
9.4 其他相关的算法	204
9.4.1 AdaCost 算法	205
9.4.2 CSB1 和 CSB2 算法	206
9.4.3 RareBoost 算法	206
9.5 重抽样的影响	207
9.6 多类别不平衡数据分类算法	211
9.7 实验分析	215
9.8 本章小结	216
第十章 基于核聚类欠抽样集成不平衡 SVM 分类算法	223
10.1 SVM 在不平衡数据下分类边界的偏移	223
10.2 基于核聚类的欠抽样 SVM 算法	225
10.2.1 传统欠抽样算法分析	225
10.2.2 基于核聚类的欠抽样	227
10.3 基于核聚类欠抽样集成 SVM 分类算法	230
10.4 核聚类欠抽样集成 SVM 算法复杂度	231
10.5 仿真实验及性能分析	232
10.5.1 实验数据	232
10.5.2 不同数据预处理算法的分类性能比较	232
10.5.3 不同比例下不平衡数据分类性能比较	234
10.5.4 不同比例下不同算法的效率比较	248
10.5.5 不同不平衡数据集分类算法性能比较	248
10.5.6 高斯核半径的参数对算法性能的影响	257
10.6 本章小结	259
第十一章 核偏移及主动学习欠抽样不平衡 SVM 算法	261
11.1 核边界偏移算法	262

11.2 KBA 算法流程	264
11.3 主动学习欠抽样策略	266
11.3.1 SVM 在不均衡数据下分类边界的偏移	266
11.3.2 传统欠抽样算法分析	267
11.3.3 基于主动学习的欠抽样	269
11.3.4 基于主动学习欠抽样不均衡 SVM 算法	270
11.4 实验分析及对比	272
11.4.1 实验数据	272
11.4.2 不同算法的分类性能比较	273
11.4.3 不同比例下不均衡数据分类性能比较	281
11.4.4 α 参数对算法性能的影响	282
11.4.5 参数 L 对算法性能的影响	291
11.5 本章小结	291
第十二章 不均衡 SVM 分类算法在故障诊断中的应用	295
12.1 故障检测的内容介绍	296
12.2 故障特征参量的选取	298
12.3 ODR-BSMOTE 故障诊断方法实验	299
12.3.1 不同比例故障数据下算法性能分析	300
12.3.2 参数 α 对算法性能的影响	301
12.3.3 参数 k 对算法性能的影响	306
12.3.4 泛化能力对比实验	307
12.3.5 分析结论	308
12.4 基于 BSMOTE 代价敏感不均衡故障诊断算法	309
12.4.1 不均衡数据对 SVM 检测性能的影响试验	309
12.4.2 不同 SVM 不均衡分类策略的对比实验	310
12.4.3 BSMOTE 参数对 SVM 性能的影响试验	313
12.4.4 泛化能力对比实验	313

12.4.5 分析与讨论	315
12.5 本章小结	315
第十三章 结论与展望	317
参考文献	323

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已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平衡样本集。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平衡数据集分类任务时往往泛化性能不高^[6]。一方面不平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]；另一方面不平衡样本给传统机器学习的研究带来了新的挑战，所以对不平衡样本的研究成为了学者们争相解决的问题。由于不平衡样本问题中各类学习样本数量不平衡，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，所以对不平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平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平衡样本

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。

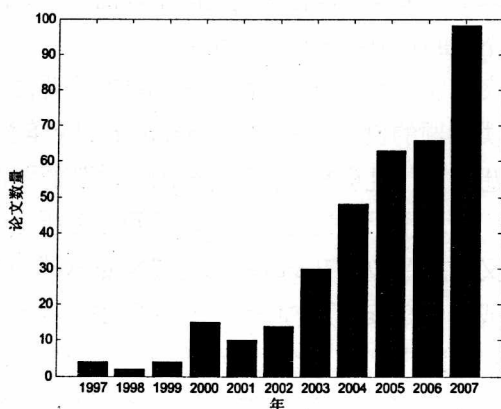


图 1.1 不均衡样本学习问题方向的研究论文发展趋势

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则，例如全局精度或是误差率，是不能给不平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，对不平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，不平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，即不平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平衡一样值得关注，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平衡。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，