统计文库

首都经济贸易大学统计前沿系列

基于重抽样的Boosting算法研究

宋捷 著

中国统计出版社
China Statistics Press

统计文库

首都经济贸易大学统计前沿系列

基于重抽样的Boosting算法研究

宋捷 著

中国统计出版社
China Statistics Press

图书在版编目(CIP)数据

基于重抽样的 Boosting 算法研究 / 宋捷著. —— 北京：中国统计出版社，2017.10

（统计文库. 首都经济贸易大学统计前沿系列）

ISBN 978－7－5037－8145－2

Ⅰ. ①基… Ⅱ. ①宋… Ⅲ. ①数理统计－研究 Ⅳ. ①O212

中国版本图书馆 CIP 数据核字(2017)第 121407 号

基于重抽样的 Boosting 算法研究

作　　者/宋　捷
责任编辑/姜　洋
封面设计/黄俊杰
出版发行/中国统计出版社
通信地址/北京市丰台区西三环南路甲 6 号　邮政编码/100073
电　　话/邮购(010)63376909　书店(010)68783171
网　　址/http://www.zgtjcbs.com
印　　刷/北京厚诚则铭印刷有限公司
经　　销/新华书店
开　　本/710×1000mm　1/16
字　　数/85 千字
印　　张/7
版　　别/2017 年 10 月第 1 版
版　　次/2017 年 10 月第 1 次印刷
定　　价/32.00 元

前　言

随着信息时代的来临，世界被大量的数据淹没了。如何从这些海量数据中找出对社会、生活有用的信息是我们面临的挑战。统计学在这样的环境中发挥着巨大的作用，因为统计学本身就是一门收集、分析、展示、解释数据的科学。按照 Breiman 的观点，统计文化包括传统建模与算法建模两种不同的文化。之所以这样划分是与统计发展中面临的不同问题分不开的。在这当中，计算机扮演了非常重要的角色。由于计算机存储能力的提高，我们所面对的数据集越来越大，数据维数越来越高，数据结构也越来越复杂。当大量的计算对我们来说已经不再是问题的时候，算法建模文化就不断地渗透到了统计的文化中，形成其中一个重要的领域：机器学习。机器学习，顾名思义，就是让机器自己学习。让机器模拟人类认知世界的过程来进行学习，从而达到能够完善自身的目的。

机器学习算法在处理海量数据上已经显示出它的优势。它与传统建模方法的根本区别在于它不需要对数据进行任何假设，而这些假设总是不能验证其是否是正确的。在处理高维数据时，比起传统建模方法它们有着非常好的抵抗维度灾难的能力。这就使得机器学习算法得到了广泛的应用与不断的发展。Boosting 算法正是近二十

年来兴起的一种机器学习算法。它是组合算法中的一种。它通过自适应地抽样生成基学习器并将这些基学习器自适应地加权组合形成一个强的学习器。自 Freund 和 Schapire 于 1995 年提出第一个 Boosting 算法——Adaboost 算法以来，Boosting 算法一直是人们研究的热点。人们对 Boosting 算法也都有自己不同的理解。根据实际问题的不同，又衍生出了许多不同的 Boosting 算法，形成了现在的 Boosting 算法族。

本书以 Boosting 算法为研究对象，首先对 Boosting 的历史、发展和研究现状进行了综述与分析。然后从实际问题出发，介绍了 7 种新的 Boosting 算法。实验结果表明，这些算法都比以前的方法有更好的有效性与适用性。本书的具体内容如下：

1. 介绍 Boosting 算法的框架、研究历程，并介绍了几种代表性的 Boosting 算法，如：Adaboost、LogitBoost、L2Boost、Acr-x4 等；

2. 当 Adaboost 算法应用到不平衡数据时会忽略少数类数据的特征。在引入类内错判率作为评价标准后，对 Adaboost 算法每一步迭代时以所有类的错判率作为迭代权重进行改进，书中介绍了两种改进的 Adaboost 算法，称为 BAboost 算法和 BAboost-J 算法。新算法与 Adaboost、Bagging、随机森林、SmoteBoost 等算法相比，在不平衡数据上能显著地降低少数类数据的分类误差。与 Adaboost 算法相比，新算法能得到更高的边际，意味着 BAboost 算法的分类能力强于 Adaboost 算法。BAboost 是指数损失下的最优解的调整。通过对

少数类数据提高权重的方法，书中还介绍了两种改进的 LogitBoost 算法。在考虑各个类内部的分类误差的情况下，讨论了 Adaboost 算法与 LogitBoost 算法的区别。新算法同样能够在不平衡数据上显著地降低少数类数据的分类误差。

3. 在回归问题中，Gradiet Boosting 算法是一种基于损失函数，是一个关于基学习器连续的函数，利用损失函数关于基学习器的一阶导数方向作为寻优方向搜索局部最优解的算法。L2Boost 算法以线性模型作为基学习器，是其典型代表。但是如果基学习器空间是不连续的，梯度 Boosting 的方法就不适用了。该类算法在理解 Adaboost 算法时，没有沿用 Adaboost 算法中的重抽样技术。这里介绍一种将重抽样技术沿用到 Boosting 回归中的新的 Boosting 回归树算法，称为 ABRT 算法。该算法能比 L2Boost 算法得到更小的预测误差，原因是该算法有着比 L2Boost 算法要小的偏差。因此，该算法牺牲了一些方差，但是得到了更小的偏差与预测误差。而该算法收敛速度比 L2Boost 要低。ABRT-1 算法则是为了提高 ABRT 算法的收敛速度而提出的一种算法。该算法预测误差更低。进一步，将基分类器的分布情况引入考虑，书中介绍了基于距离的 Boosting 算法。该算法收敛速度也比 ABRT 要快，而且预测误差也同样很小。

本书是在导师吴喜之教授的悉心指导下完成的。感谢吴老师和师母沈老师在学习和生活等各个方面对我的悉心教导与亲切关怀。其次，我还要感谢中国人民大学统计学院的所有老师，特别是易丹辉

老师和张波老师，他们在我博士学习期间对我的关怀与指导，让我受益匪浅。还要衷心地感谢师姐吕晓玲、闫洁，师兄马景义、魏传华、田茂再、陈凯、马国栋的无私帮助。还要感谢赵秀丽同学和刘苗师妹的帮助。

本书的出版也得到了首都经济贸易大学统计学院的支持和帮助。学院的领导和同事们都给了我各个方面的启发和帮助。本书还得到“国家自然科学基金青年基金(11201316)”的资助。最后我要向我的家人表示由衷的感谢，是他们的支持与鼓励使我走到现在。

宋捷谨记

2017 年 7 月于北京

目　录

第1章
绪 论

1.1 选题意义

Boosting算法是近二十年来兴起的一种机器学习[①]算法。它是集成算法中的一种。但是它和其他的集成算法，如：Bagging、随机森林是完全不同的。它不只是简单地集成。它最大的优点在于它的自适应性。由于它可以有效地减小偏差和方差，现在已经被广泛地应用到了许多领域。由于Boosting在每次更新时对误差大的样品自适应地赋予了更大的权重，故能够有效地减小预测误差。目前流行的观点是由Friedman等人提出的，认为Adaboost是一种基于指数损失的Logistic可加模型的拟牛顿迭代算法。并且由Breiman、Jiang等人已经证明了在指数损失下的Adaboost算法的一致性。但是基于不同的损失与不同的组合方式，研究者们诸如Friedman、Mason又提出了很多不同的Boosting算法，而且这些新的算法同样能得到与Adaboost算法相比的预测误差。也就是说Breiman、Friedman等人对Boosting算法的解释是不够的。Boosting算法的运行机制到现在还是没有解决的问题。因此，研究在广义的一个损失下的Boosting算法的性质是非常有意义的。对Boosting算法的运行机

① (Dietterich，1997)一文中提出了机器学习中的四个研究方向。

制提出新的理解也是十分重要的。

分类和回归是统计应用中的两个重要领域。Boosting 算法作为一种提升的机器学习算法，在这两个领域已经得到了越来越广泛的应用。本书将从 Boosting 算法在这两个领域的应用入手，讨论其运行机制，对 Boosting 算法做出一种新的解释。

分类通常是通过训练数据学习得到一个分类器，然后再用这个分类器对需要预测的数据进行预测。目的是得到较好的预测，也就是得到更小的预测误差。在以错分率作为评价标准时，往往会造成选出来的错分率最小的分类器并不是想要的好的分类器。实际中会遇到很多这样的数据。比如在诊断病人时，我们有的只是很少的病例，而绝大多数人都是健康人。这时病人和健康人就是比例相差很大的两个类。诊断病人时总是希望把病人都诊断为病人，健康人都诊断为健康人。但是不管怎样都会有错分发生，而且把病人诊断为健康人的风险总是远远高于把健康人诊断为病人。也就是说目的是使得两个人群中各自的错分都要比较小，尤其是病人人群中的错分率。可见，此时的评价标准不再是总的错分率，而是各个类内部的错分率。因此，如何能够降低模型在少数类数据上的分类误差将是很有意义的事情。

Adaboost 算法是一种预测误差很低的分类算法。因为其良好的分类能力与抵抗过拟合的能力而受到大家的广泛应用。Breiman 用 boosting 技术来提高一次分类的能力，通过集成的方法有效地减小偏差，从而提高单个不稳定分类的预测误差。Adaboost 算法的运行机制也一直是大家研究的热点。Adaboost 算法通过自适应地迭代优化的方法来进行分类。Breiman、Zhang 等人指出，在数据对称的条件下，Adaboost 算法的预测误差在样本足够大的情况下的极限是贝叶斯误差。由于在不平衡数据下，分类误差已经不能作为模型的评价标准，那么即便是更接近贝叶斯误差也不能说明这时的模型是更优的。因此，将 Adaboost 改进以适用于不平衡的数据分类是非常有意义的工作。

Friedman 等人将 Adaboost 算法看作是对 Logistic 可加模型在各种损失下的拟合算法。在负二项损失下，他们提出了一种新的算法 LogitBoost 算法。这个算法与 Adaboost 有着相当的分类能力。但是它与 Adaboost 一样，在不平

衡数据上也会带来非常高的少数类数据的错分率。因此，将 LogitBoost 改进以适用于不平衡数据分类也是很有意义的。另外，对 Adaboost 和 LogitBoost 两种不同的算法而言，两者不同的根本仍是还没有解决的问题。这里借助不平衡数据，讨论 Adaboost 和 LogitBoost 的不同之处也是一个新的研究方向。

直到现在，boosting 技术在回归中的讨论远远没有像分类中讨论得那样多。也没有像 Bagging 那样在回归中讨论得多。究竟 boosting 在回归中是降低了偏差还是降低了方差，还是对某种类型的数据来说起到降低均方误差的作用，这些都还是没有解决的问题。所以研究 Boosting 算法在回归中的作用是非常重要的。而且目前 Boosting 算法的研究主要是在连续泛函空间中的研究，特别是基于线性模型的 Boosting 回归。当选择树这样不连续的回归器作为基学习器时，实际上选择的寻优的基空间将是由一些阶梯函数构成的空间。与此同时，对总体分布的假设不再有连续分布函数的假设。整个基学习器空间是不连续的，那么现有的梯度 Boosting 的思想来解释 Boosting 算法就不适用了。所以研究不连续空间下的 Boosting 算法也是很有意义的。

1.2 本书主要内容

本书的主要内容包括以下两个大的方面：

第一部分：讨论 Boosting 算法在分类中的若干问题

1)在这部分中，首先讨论 Adaboost 算法在处理不平衡数据中的不足，然后介绍一类改进的 Adaboost 算法：BAboost 算法。这类算法在引入类内错判率作为评价标准后，对 Adaboost 算法每一步迭代时以所有类的错判率作为迭代权重进行了改进。该算法包括了适用于两分类与多分类的两种新算法。新的算法在每一步迭代时以每个类的类内错判率作为更新权重。BAboost 算法是在指数损失下对最优解的一个调整。实验结果表明，新算法的边际比 Adaboost 算法要高，说明该算法在不平衡数据上的分类能力比 Adaboost 要好。并且利用模拟数据讨论了新算法中的参数与各个类内部的预测误差之间的关系。利用实际数据将新算法与 Adaboost 算法、SMOTEBoost 算法、

Bagging 算法、决策树和随机森林进行比较。利用实验讨论新算法中参数的一般取值。从 BAboost 算法看来，对数据给予不同的权重，考虑基分类器的不同分布，都会影响预测误差。

2）LogitBoost 算法在处理不平衡数据时与 Adaboost 算法有着同样的不足，一种改进的 LogitBoost 算法：BLogitBoost 算法，是在通过增加少数类数据权重的思想下形成的。其中也包括适用于两分类与多分类的两种新算法。新的算法在下一步的迭代中都增加了少数类数据的权重。利用模拟数据实验，讨论了新算法中的参数与各个类内部的预测误差之间的关系。利用实际数据将几个新算法与 Adaboost 算法、LogitBoost 算法进行了比较。从 BLogitBoost 算法看来，同样的，对数据采用不同的抽样权重是会影响预测误差的。

第二部分：讨论 Boosting 算法在回归中的若干问题

这部分主要讨论基于重抽样技术的 Boosting 算法在回归中的性质。Gradiet Boosting 算法的特点是损失函数关于基学习器连续，并利用损失函数关于基学习器的一阶导数方向作为寻优方向搜索局部最优解的算法。L2Boost 算法以线性模型作为基学习器，是其典型代表。但是如果基学习器空间是不连续的，梯度 Boosting 的方法就不适用了。该类算法在理解 Adaboost 算法时，没有沿用 Adaboost 算法中的重抽样技术。这里将重抽样技术沿用到 Boosting 回归中所提出的一种新 Boosting 回归树算法——ABRT 算法。实验结果表明，新算法没有 L2Boost 算法收敛快，但是预测误差总是比 L2Boost 算法要小。数据可加性越弱，该算法越容易发生过拟合，因为算法是以可加模型为基础的。该算法比起 L2Boost 算法降低了偏差，因为 L2Boost 算法的方差很小，几乎为 0，故其偏差非常大。该算法牺牲了一些方差，但是得到了更低的预测误差。另外，文中还对不同的重抽样概率进行了讨论。发现对不同的数据来说，不同的重抽样概率将会导致不同的预测误差。书中还介绍了一种为了提高 ABRT 算法的运算速度而提出的一种改进的 ABRT 算法——ABRT-1 算法。该算法在提高运算速度的同时，并没有增加预测误差。将基分类器的分布引入，书中又介绍了一种基于距离的重抽样 Boosting 回归方法。实验表明，该算法大大提高了 ABRT 算法的收敛速度，预测误差比之稍微增加，但还是小于

L2Boost 算法。

1.3　本书的结构安排

本书分为五章。第 1 章即为本章，是绪论部分。第 2 章是对 Boosting 算法进行综述，介绍了 Boosting 算法的流程、Boosting 算法的研究历程、Boosting 的代表算法以及 Boosting 算法的应用。第 3 章在对不平衡数据的研究现状进行综述之后，介绍了四种解决不平衡数据问题的新算法：BAboost 算法、BAboosT-J 算法、BLogitBoost 算法、BLogitBoosT-J 算法。然后对新算法的运行机制与 Adaboost、LogitBoost 的运行机制的不同进行了解释和比较。模拟数据与实际数据的结果表明，新算法能够显著地降低少数类的预测误差。第 4 章在对 Boosting 算法在回归中应用的研究现状进行综述之后，介绍了三种新的基于重抽样回归树的 Boosting 回归算法：ABRT、ABRT-1 与 ABRT-D 算法。并且证明了前者在训练集上的收敛性，最后以最近邻的思想来解释 Boosting 算法的运行机制。模拟数据与实际数据的结果表明，新算法也都能得到更低的预测误差，并且后两者的运算速度更快。第 5 章是结论与对未来工作的展望。此章是对全书的一个总结，提出作者对 Boosting 算法运行机制的一种理解。这在上述所提到的几个新算法中都有体现，而且实验表明，Boosting 算法在这种思想理解下使得 Boosting 算法族更为广泛，更有实用性，效果也更好。

第 2 章
Boosting 算法综述

2.1 引言

在计算机科学领域，现在大家普遍认同 Boosting 算法的思想来源于由 Valiant 于 1984 年提出的 PAC(Probably Approximately Correct)(Valiant L. G. ,1984)学习模型。在提出这个学习模型的同时，Valiant 和 Kearns 提出了弱分类器与强分类器的概念。弱分类器[①]是指预测精度稍稍高于随机猜想的分类器，也就是说预测精度稍大于 1/2。强分类器是指预测精度很高并能在多项式时间内完成的分类器。并且，他们首次提出 PAC 学习模型中弱分类器与强分类器的等价性问题，即：任意给定弱分类器，能不能将其提升为强分类器。如果回答是肯定的，那么只需找到一个弱分类器就可以将其提升为一个强分类器，因此就不必再很难地寻找强学习器。

Schapire 于 1990 年最先构造出一种多项式级的算法(Schapire R. E. , 1990)，对该问题做了肯定的回答，这就是最初的 Boosting 算法。到了 1996 年，Freund 和 Schapire 通过改进之前的 Boosting 算法，提出了现在大家经常用到的 Adaboost 算法(Freund Y. ,1996)。在此以后，人们不断探求其背后真

① (Freund,1995)文献中对弱学习器进行了介绍。

正的原理，又提出了许多基于 Boosting 思想的很多新的 Boosting 算法，丰富了 Boosting 的内容。

统计学家一直致力于找出 Boosting 的工作原理。最早做这个尝试的是 Leo Breiman。Breiman 认为，Adaboost 算法是 Arcing 算法中的一种(Breiman L.，1998)。紧接着，Schapire 与 Freund 等人(Schapire et al.，1998)提出 Adaboost 算法的本质是提升了单个弱分类器在各个类之间的边际。后来，Friedman 等人(Friedman et al.，2000)则认为 Boosting 算法是可加模型的一种梯度寻优算法。Breiman (Breiman L.，2000)、Zhang (Zhang T.，2004，2005)、Jiang(Jiang W.，2004)等人也对 Booosting 算法的一致性进行了探讨。Mease 和 Wyner 等人(Mease D.，Wyner A.，2008)又对 Friedman 等人的观点提出质疑。本章将就这些方面做出一个 Boosting 算法的综述。

2.2　Boosting 算法描述

Boosting 算法是一种组合算法，由训练集所生成的子集来建立基学习器，然后再将这些基学习器整合起来。实际上，Boosting 算法是一个算法族。它是一种思想，同时也是一种框架。对这个框架的不同的理解可以延伸出不同的算法。首先我们先来看看 Boosting 的框架。以 A 表示原始的训练集，以 $(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)$ 表示其中一个观测数为 n 的训练集。在这个子训练集上训练得到一个估计函数记为 $\hat{g}(\cdot)$，它就是基学习器。然后再将这些函数组合起来得到一个最终的估计，也就是说将这些基学习器组合起来。其工作原理如图 2—1 所示。

这里加权的数据集就是在原始数据训练集上分别给 n 个样品以一定的权重所生成的子训练集。简单地，可以认为这时的基学习器是一个加权的估计函数，比如加权最小二乘估计等。一般来说这些加权的数据集的权重都是不一样的。但是这也与我们对这个框架的理解的不同而不同。最后整合后的估计 $\hat{f}(\cdot)$ 则是基学习器的一个加权组合，权重是 α_m。

从图 2—1 的框架看来，对 Boosting 算法而言，最重要的问题有三个。一

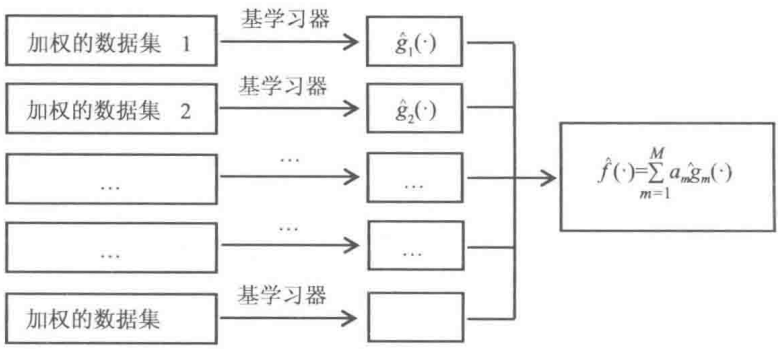

图 2－1　Boosting 算法的工作原理图

个是对原始数据集如何加权的问题。二是基学习器选择什么基学习器的问题，比如决策树、支持向量机、神经网络等①。三是对基学习器如何加权组合的问题。本书只研究以决策树作为基学习器的 Boosting 算法。因此问题就变成了两个。

提到如何对数据进行加权，能最先想到的就是加权最小二乘方法。这是一种最简单的加权方法。当数据有异方差的状况时，利用标准差对数据进行加权将会得到很好的性质。假设模型

$$Y=f(X)+\varepsilon=\beta X+\varepsilon,\varepsilon\sim N(0,\sigma^2(X))$$

对参数 β 来说，这时 $(X'X)^{-1}X'Y$ 不再是其线性最小方差无偏估计。如果对模型进行变换：

$$Y/\sigma(X)=\beta X/\sigma(X)+\varepsilon/\sigma(X)，\text{记为 } Y^*=\beta X^*+\varepsilon^*$$

这时新模型下的 β 的估计 $(X^{*\prime}X^*)^{-1}X^{*\prime}Y^*$ 将是其线性最小方差无偏估计。这种加权方法是用数据的分布的参数形式来对数据进行加权。相应地，还有用数据的分布的非参数形式来对数据进行加权的方法。这个在后面再详细讨论。

对数据加权的问题还与给出的损失函数 L(f(X),Y)有关。比如，考虑平方损失 $L(f(X),Y)=E(Y-f(X))^2$ 与绝对值损失 $L(f(X),Y)=E|Y-f(X)|$ 的情况。如图 2－2，当采用平方损失时，实际上赋给了绝对值大的数更高的权

① 这些基学习器在很多书中都不同侧重地有详细的介绍。比如参考文献：[1]，[2]，[3]，[6]，[7]，[10]，[53]等。