第32卷 第3期
2017年6月

北京信息科技大学学报
Journal of Beijing Information Science& Technology University

Vol. 32 No. 3
Jun. 2017

文章编号:1674-6864(2017)03-0049-05

DOI: 10.16508/j.cnki.11-5866/n.2017.03.010

基于 AdaBoost 的二手车价值评估方法

刘 聪,程希明

(北京信息科技大学 理学院,北京 100192)

摘 要: 将自适应提升方法(AdaBoost)应用于二手车价值的评估,提出一种以决策树桩作为弱分类器的集成方法。二手车数据样本量大,实体特征多,通过区间离散化得到样本集,避免深度遍历产生过拟合。通过加权表决集成弱分类器,建立分类模型。实验表明,AdaBoost 方法相比传统的决策树方法,准确率提高 7.1%。

关 键 词: 二手车评估; AdaBoost; 二类分类; 决策树桩; 分类器系数

中图分类号: TP 181 **文献标志码:** A

Evaluation method of used car value based on AdaBoost

LIU Cong, CHENG Ximing

(School of Applied Science, Beijing Information Science & Technology University, Beijing 100192, China)

Abstract: To apply adaptive boosting (AdaBoost) to the evaluation of used car value, a boosting method of decision stump as a weak classifier is proposed. Because of the large quantities of sample data and features, the sample set is got through interval discretization to avoid the over-fitting due to the depth of traversal. The weak classifier is integrated by weighted voting to establish the classification model. The experiments prove that the accuracy of the AdaBoost is 7.1% higher than that of the traditional decision tree.

Key words: used car evaluation; AdaBoost; binary classification; decision stump; classifier weight

0 引言

目前,按国家规定的二手车评估方法有重置成本法、收益现值法、现行市价法和清算价格法[1]。针对二手车价值评估问题的数据挖掘方法有聚类分析[2]、因子分析[3]、决策树算法[4]、支持向量机[5]等。这些方法从不同角度对二手车价值进行评估分析,但都有其局限性。因子分析通过计算公因子的贡献率,得到影响二手车价值的因素,但不能对二手车价值做出实际预测[3]。决策树算法以树的形式表现分类规则,在处理大数据集时,会增加分类算法的计算复杂性[6]。支持向量机(Support Vector Machine,SVM)对于小样本数据的分类准确率高,有较好的知识泛化能力,但 SVM 求解是一个凸二次规划的过程,当样本容量过大时,时间复杂度较大[6]。AdaBoost 方法是一种集成的学习方法[7],它能够将学习的正确率仅比随机猜测略好的弱学习器提升为学习正确率高的强学习器,从而有效提高学习精度。

二手车数据样本量大,实体因子对价值影响明显,实体特征较多。AdaBoost 方法具有高精度、方法简单无需做特征筛选,且不会过度拟合的优点,适用于二手车价值的分类。本文以决策树桩作为弱分类器的 AdaBoost 提升方法,以信息增益率构建决策树桩,通过改变决策树桩权值,弥补了决策树方法由于深度遍历而缺乏伸缩性的缺点。

1 基于 AdaBoost 的二手车价评估方法模型

1.1 数据准备

本文使用网络爬虫,抓取国内 O2O 二手车服务商(瓜子二手车、人人车等)网站上展示的车辆信息,作为样本数据。截止 2016 年 10 月,抓取出网站

收稿日期: 2017-01-09

基金项目: 国家自然科学基金资助项目(9011323905)

作者简介: 刘 聪,女,硕士研究生;通讯作者:程希明,男,博士,教授。

在售二手车数据共39 399条。通过数据清洗,得到规范化数据32 859条。价值合理车辆数据10 928条,价值不合理车辆数据21 931条。本文默认数据集中车辆信息与实际车辆状况一致。

根据文献[3],确定以车型、变速箱、使用时间、折旧率、排量、行驶里程为候选属性,以二手车价格合理类别(是/否)为目标属性。由于二手车数据中排量、上牌时间、行驶里程、新车价、新旧车差价为连续型数值,且各属性数据单位和值域均不相同,因此需对其连续性数据离散化。按照国家车辆分类标准,将"车型"属性分为9类,将"变速箱"分为6类,将"排量"分为7类,将"行驶里程"分为7类。特殊地,"使用时间"定义为当前时间减去"上牌时间",并将其离散化为11类,折旧率为新旧车差价与新车价比值,离散化为5类。数据属性取值详细描述如表1所示,部分二手车训练样本如表2所示。

表1 数据属性及取值

属性	描述	属性值	分类数
A_1	车型	A00, A0, A, B, C, D, E, SUV, OTHER	9
A_2	变速箱	AT, MT, TIPT, CVT, DC, AMT	6
A_3	排量/T	<1.0, 1.1~1.6, 1.7~2.0, 2.1~2.5, 2.6~3.0, 3.0~4.0, >4.0	7
A_4	行驶里程/km	<1, 1~3, 3~5, 5~8, 8~10, 10~15, >15	7
A_5	使用时间/年	<1, 1, 2, 3, 4, 5, 6, 7, 8, 9, >9	11
A_6	折旧率	<0.2, 0.2~0.4, 0.4~0.6, 0.6~0.8, >0.8	5
Y	是否合理	0, 1	2

表2 部分二手车数据样例

A_1	A_2	A_3	A_4	A_5	A_6	Y
A0	AT	1.1~1.6	5~8	7	0.2~0.4	1
A	MT	1.1~1.6	8~10	7	0.2~0.4	0
SUV	AT	2.6~3.0	8~10	4	0.6~0.8	1
SUV	TIPT	3.0~4.0	3~5	2	0.6~0.8	1
E	TIPT	1.7~2.0	1~3	3	0.6~0.8	1
A	MT	1.1~1.6	10~15	6	0.4~0.6	0
B	TIPT	1.7~2.0	5~8	2	0.6~0.8	1
A	TIPT	1.7~2.0	5~8	6	0.4~0.6	0
B	TIPT	1.7~2.0	8~10	6	0.4~0.6	0
OTHER	TIPT	2.1~2.5	1~3	1	>0.8	0
SUV	TIPT	1.1~1.6	3~5	1	0.6~0.8	0
A	TIPT	1.7~2.0	5~8	4	0.4~0.6	0
C	CVT	1.7~2.0	8~10	6	0.4~0.6	0
A	MT	1.1~1.6	3~5	6	0.4~0.6	1
SUV	TIPT	1.7~2.0	10~15	6	0.4~0.6	0
C	TIPT	1.7~2.0	5~8	3	0.6~0.8	1
B	MT	1.7~2.0	5~8	5	0.4~0.6	0
A0	MT	1.1~1.6	10~15	7	0.4~0.6	1
B	TIPT	1.7~2.0	3~5	2	0.6~0.8	0
A	AT	1.1~1.6	8~10	8	0.2~0.4	0
SUV	CVT	2.1~2.5	10~15	6	0.2~0.4	0

1.2 建立模型

二手车数据量较大,实体特征较多,且特征之间不相互独立。AdaBoost方法的优点是,方法简单,不需要做特征筛选,高精度,且不会过度拟合。同时,决策树桩不需要进行深度搜索。因此本文采用以决策树桩作为弱分类器的AdaBoost方法建模。

首先定义弱分类器,弱分类器由 $x = v$ 或 $x \neq v$ 产生,即

$$G(x) = \begin{cases} 1 & x = v \\ 0 & x \neq v \end{cases} \tag{1}$$

式中,v 为二手车某个属性值。

设二手车数据集为 D,$|D|$ 为二手车数据容量。二手车价值合理性有2类,表示为 $Y_k(k = 0,1)$,$|Y_k|$ 为属于 Y_k 类别的二手车个数。二手车数据属性为 $A_j(j = 1,2,\cdots,6)$,每个属性的分类个数如表1所示。设属性 A_j 分类个数为 n,根据属性 A_j 的取值,将 D 划分为 n 个子集 $D_{j1},D_{j2},\cdots,D_{jn}$,$|D_{ji}|$ 为 D_{ji} 的样本个数。

计算每个属性结点的信息增益 $g(D,A_j)$,选择信息增益最大的属性作为弱分类器决策树桩根结点。属性 A_j 对数据集 D 的信息增益为

$$g(D,A_j) = H(D) - H(D \mid A_j) \tag{2}$$

式中,$H(D)$ 为二手车训练集 D 的经验熵:

$$H(D) = -\sum_{k=0}^{1} \frac{|Y_k|}{|D|} \log_2 \frac{|Y_k|}{|D|} \tag{3}$$

$H(D|A_j)$ 为属性 A_j 给定条件下 D 的经验条件熵:

$$H(D \mid A_j) = \sum_{i=1}^{n} \frac{|D_{ji}|}{|D|} H(D_{ji}) = -\sum_{i=1}^{n} \frac{|D_{ji}|}{|D|} \sum_{k=0}^{1} \frac{|Y_{ik}|}{|D_{ji}|} \log_2 \frac{|Y_{ik}|}{|D_{ji}|} \tag{4}$$

信息增益最大的属性结点 A_j 作为根结点的属性。由于 A_j 有 n 个属性值,由式(1)可知弱分类器定义,然后计算每个属性值的分类误差率 $e_i(i = 1,2,\cdots,n)$,找到最低误差率的属性值 v。

其次,建立二手车价值评估的AdaBoost模型,对弱分类器加权表决分类。描述如下:

已知二手车数据集 $D = \{(x_1,y_1),\cdots,(x_N,y_N)\}$,$x_i$ 属于 n 维实例空间,y_i 属于标记集合 $\{0,1\}$。

算法步骤:

输入:二手车数据集 D,弱学习算法 P。

输出:强分类器 $G(x)$

Step1 初始化权值 $\omega_i = \frac{1}{N}(i = 1,2,\cdots,N)$;

Step2　DO FOR $m = 1,2,\cdots,M$(m 为第 m 轮训练)

1) 根据弱学习算法 P, 由式(1) 可得弱分类器;

2) 计算 $G_m(x)$ 在训练数据集上的分类错误率:

$$e_m = \sum_{i=1}^{N} w_{mi} I(G_m(x_i) \neq y_i) \tag{5}$$

3) 计算弱分类器系数:

$$\alpha_m = \frac{1}{2}\ln\frac{1-e_m}{e_m} \tag{6}$$

4) 调整样本权值:

$$\omega_{m+1,i} = \frac{\omega_{mi}}{Z_m}\exp(-\alpha_m y_i G_m(x_i)) \tag{7}$$

式中, $i = 1,2,\cdots,N$; Z_m 为归一化因子。

$$Z_m = \sum_{i=1}^{N} \omega_{mi}\exp(-\alpha_m y_i G_m(x_i)) \tag{8}$$

5) 判断循环终止条件:

如果 $\alpha_m \leqslant \varepsilon$, ε 为弱分类器系数的阈值, 则令 $M = m$, 跳出循环。

END

Step3　将弱分类器集合成强分类器:

$$G(x) = \mathrm{sign}(f(x)) \tag{9}$$

其中 $f(x)$ 为基本分类器线性组合:

$$f(x) = \sum_{m=1}^{M} \alpha_m G_m(x) \tag{10}$$

AdaBoost 算法的加权表决, 将被弱分类器错分的样本权值增加, 正确分类的样本权值缩小, 根据加权错误率最小来选取新的弱分类器。

利用本文 1.1 节的二手车价值合理性数据集 D, 建立二手车价值合理性预测模型。详细步骤如下:

1) 初始化数据的权值分布。初始权值为 $\omega_1 = \frac{1}{32\ 859}$, 其中 32 859 为二手车数据样本容量;

2) 在权值分布为 ω_1 的训练数据集 D_1 上, 根据信息增益准则选取最优属性为根结点, 作为决策树桩进行二分类;

计算各个属性的信息增益, 选取信息增益最大的属性, 得到弱分类器 $G_1(x) = \begin{cases} 1 & A_4 = ''<1'' \\ 0 & A_4 \neq ''<1'' \end{cases}$

3) 计算 $G_1(x)$ 的错误率和 $G_1(x)$ 的系数。由式(5) , $G_1(x)$ 对 D_1 进行分类的错误率为

$$e_1 = \sum_{i=1}^{32\ 859} \omega_1 I(G_1(x_i) \neq y_i) = 0.332\ 6$$

由式(6) , 计算弱分类器 $G_1(x)$ 的系数

$$\alpha_1 = \frac{1}{2}\ln\frac{1-e_1}{e_1} = 0.348\ 2$$

4) 调整样本权值 ω_2

初始样本权值为 ω_1, 当执行步骤 2) 、步骤 3) 后, 根据式(7) 调整样本权值, 使下一轮分类的样本数据以调整后的样本权值 ω_2 更新数据 D_2, 以 D_2 为新的数据重新执行步骤 2) 、3) 、4) 。

循环执行步骤 2) 、3) 、4) 直到 D_{10} (设置弱分类器数目为 10, 能够得到 10 个弱分类器 $D_1, D_2, \cdots, D_{10}$ 及其系数 $\alpha_1, \alpha_2, \cdots, \alpha_{10}$。如表 3 所示) 。

表 3　弱分类器及其系数

$G_m(x)$	弱分类器系数
$G_1(x) = \begin{cases} 1 & A_4 = ''<1'' \\ 0 & A_4 \neq ''<1'' \end{cases}$	0.348 2
$G_2(x) = \begin{cases} 1 & A_4 = ''1 \sim 3'' \\ 0 & A_4 \neq ''1 \sim 3'' \end{cases}$	0.752 3
$G_3(x) = \begin{cases} 1 & A_5 = ''<1'' \\ 0 & A_5 \neq ''<1'' \end{cases}$	0.844 3
$G_4(x) = \begin{cases} 1 & A_6 = ''0.6 \sim 0.8'' \\ 0 & A_6 \neq ''0.6 \sim 0.8'' \end{cases}$	0.409 6
$G_5(x) = \begin{cases} 1 & A_5 = ''3'' \\ 0 & A_5 \neq ''3'' \end{cases}$	0.219 7
$G_6(x) = \begin{cases} 1 & A_5 = ''6'' \\ 0 & A_5 \neq ''6'' \end{cases}$	0.242 3
$G_7(x) = \begin{cases} 1 & A_3 = ''1.7 \sim 2.0'' \\ 0 & A_3 \neq ''1.7 \sim 2.0'' \end{cases}$	0.153 2
$G_8(x) = \begin{cases} 1 & A_6 = ''0.4 \sim 0.6'' \\ 0 & A_6 \neq ''0.4 \sim 0.6'' \end{cases}$	0.173 9
$G_9(x) = \begin{cases} 1 & A_5 = ''1'' \\ 0 & A_5 \neq ''1'' \end{cases}$	0.062 2
$G_{10}(x) = \begin{cases} 1 & A_4 = ''<1'' \\ 0 & A_4 \neq ''<1'' \end{cases}$	0.103 7

5) 分类器的确定

根据加权表决的思想, 根据式(9) 将每轮基本分类器及其系数加权组合, 就得到了 AdaBoost 二手车价值合理性分类模型。

1.3 验证和模型评价

考虑模型的预测准确率和时间开销，若准确率有显著提高，并且计算时间显著下降，说明该模型可行。

定义覆盖率和命中率作为二手车价值合理性模型评价指标。覆盖率(Recall)定义为实际价值合理，预测价值也合理的二手车数量占全部实际二手车价值合理数量的比例，计算公式为 $\frac{T_p}{T_p + F_p} \times 100\%$；命中率(Precision)定义为实际价值合理、预测价值也合理的二手车数量占全部预测价值合理的二手车数量的比例，即，$\frac{T_p}{T_p + F_n} \times 100\%$。式中，$T_p$ 为实际价值合理、预测价值也合理的二手车数量，F_p 为实际价值合理而预测价值不合理的二手车数量，F_n 为实际价值不合理但预测价值合理的二手车数量。此外，F 值(F-Measure)、接受者操作特征曲线下面积(Receiver Operating Characteristic Area，ROC Area，简称ROC面积)也能够作为模型准确率的评价参考指标。F 值为覆盖率和命中率的调和均值，F 值越接近1模型准确率越高；ROC面积越接近1，说明分类效果越好。

1.4 模型预测结果分析

实验使用WEKA平台的AdaBoost算法(weka.classifier.meta.AdaBoostM1)作为分类器，弱分类器为决策树桩(weka.Classifiers.trees.DecisionStump)。对二手车数据建立预测模型，并对预测结果进行分析。

二手车数据集样本共有32 859条，从中分别随机选择5%、10%、20%、40%、50%、60%、70%的二手车数据作为AdaBoost算法训练数据集，对应的剩余二手车数据作为测试数据集，在相同设备和平台下进行多次预测。不同数量训练集预测准确率如表4所示。以训练数据为70%、测试数据为30%为例，对二手车数据价值合理性进行预测，预测结果见表5。

表4 不同数量训练集预测准确率

训练样例占比/%	分类正确个数	分类错误个数	预测准确率/%
5	24 661	6 555	79.001 2
10	24 245	5 328	81.983 6
20	22 090	4 017	84.613 3
40	17 061	2 655	86.533 8
50	14 168	2 261	86.237 8
60	11 533	1 611	87.743 5
70	8 465	1 393	85.869 3

表5 AdaBoost建模预测结果

类别	真阳率	假阳率	命中率	覆盖率	F 值	ROC面积
0	0.870	0.148	0.774	0.870	0.819	0.796
1	0.852	0.130	0.918	0.852	0.884	0.796
各类别加权平均	0.859	0.138	0.834	0.859	0.841	0.796

从表4可看出，当训练数据达到10%以上时，预测准确率达80%以上，且训练数据集比例越大，其预测准确率越高，基本可以保持在86%左右。

由表5可知，预测结果中覆盖率为85.9%，命中率为83.4%，覆盖率和命中率均在80%以上，说明模型预测准确率较高。此外，表5中 F 值为0.841，同样在80%以上；ROC面积为0.796，在0.7~0.9范围内，有一定的准确度。以上结果均能说明该模型较高的预测准确率。

2 对比分析

为验证本模型的性能，进行对比分析实验。分别运用传统的决策树方法和AdaBoost算法对二手车数据建模。

选取2016年5月份的2837条全国二手车价值合理性数据集，分别训练得到AdaBoost模型和传统决策树模型。选取2016年6月份的2591条记录作为验证集，分别对AdaBoost模型和传统决策树模型验证，得到2种模型的验证结果。2种模型的比较结果如表6所示。

表6 比较结果

模型	训练结果		验证结果	
	覆盖率/%	命中率/%	覆盖率/%	命中率/%
AdaBoost	84.9	83.4	81.9	81.4
决策树	79.2	77.9	74.8	74.3

表6所示结果表明，使用AdaBoost方法训练的模型，验证结果比训练结果的覆盖率与命中率稍差，但整体比传统的决策树算法模型的覆盖率与命中率高。AdaBoost模型相比传统决策树模型，验证结果准确率提高7.1%。由于建模规则和分类的时间也是模型评判的一个重要指标，本实验中AdaBoost算法与决策树算法的训练时间之比为0.93，分类时间的比值为0.87，说明AdaBoost方法与决策树算法的训练时间和验证时间基本接近。

3 结束语

使用数据挖掘方法对分类问题进行预测是近年来分类领域的热点研究方向。AdaBoost提升方法是

分类算法中容易实现、精度较高的分类方法。对于二手车价值合理性评估问题,本文通过抓取网站数据作为样本数据,将其离散化,建立了以决策树桩为弱分类器的 AdaBoost 分类模型。将 AdaBoost 分类模型与传统决策树模型进行比较,发现验证结果准确率提高 7.1%。实验证明,AdaBoost 方法对二手车价值合理性评估问题具有明显优势。但是,由于地域不同、消费水平不同,评估师对二手车的评测效果也有所不同,在后续的工作中,将考虑地域差异性,分别针对不同消费习惯的地域做进一步的二手车评估工作。

参考文献:

[1] 周遊. 适用于二手车的价值评估方法研究[J]. 黑龙江交通科技,2013(11):172 -174.

[2] 潘俊. 基于聚类分析的二手车成新率计算[D]. 上海: 上海交通大学,2007.

[3] 冯秀荣,王斌. 影响二手车价值的因子分析[J]. 商业研究,2008(2):102 -105.

[4] 周凌云. 决策树在汽车评测中的应用研究[J]. 中南民族大学学报,2012,31(3):97 -100.

[5] 邱海波,钱忠民,钱默抒. 合成少数类过采样过滤器方法在二手车推荐中的应用[J]. 计算机与现代化,2016(7):118 -123.

[6] Jiawei Han,Micheline Kamber. 数据挖掘概念与技术[M]. 范明,孟小峰. 译. 北京: 机械工业出版社,2006:198 -224.

[7] 曹莹,苗启广,刘家辰,等. AdaBoost 算法研究进展与展望[J]. 自动化学报,2013,6(39):745 -758.

(上接第 48 页)

$\omega \in L^p[0,1]$:$p > 1$,$p = 1$ 和 $p = \infty$ 三种情况,证明了边值问题式(1)在满足新的充分条件时存在正解的唯一性。从方法和结果 2 个方面改进和推广了文献[1]的相关工作。

参考文献:

[1] Zhang X,Ge W. Symmetric positive solutions of boundary value problems with integral boundary conditions[J]. Applied Mathematics & Computation,2012,219:3553 -3564.

[2] Zhang X,Feng M. Green's function and positive solutions for a second-order singular boundary value problem with integral boundary conditions and a delayed argument[J]. Abstract and Applied Analysis, 2014(9):393187.

[3] Zhang X,Feng M. Positive solutions for a second-order differential equation with integral boundary conditions and deviating arguments[J]. Boundary Value Problems,2015(1):1 -21.

[4] Feng M,Ji D,Ge W. Positive solutions for a class of boundary-value problem with integral boundary conditions in Banach spaces[J]. Journal of Computational & Applied Mathematics,2008,222(2):351 -363.

[5] Zhang X,Feng M,Ge W. Existence results for nonlinear boundary-value problems with integral boundary conditions in Banach spaces[J]. Nonlinear Analysis, 2008, 69(10): 3310 -3321.

[6] Zhang X,Feng M,Ge W. Existence result of second-order differential equations with integral boundary conditions at resonance[J]. Journal of Mathematical Analysis & Applications, 2009,353(1):311 -319.