计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2018,27(2):202−206 [doi: 10.15888/j.cnki.csa.006203]
©中国科学院软件研究所版权所有.
E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于 Adaboost 的随机森林算法在医疗销售预测中的应用①

常晓花, 熊　翱

(北京邮电大学 网络技术研究院, 北京 100876)
通讯作者: 常晓花, E-mail: vanessachang8@outlook.com

摘　要: 提出一种基于 Adaboost 方法的随机森林销售量预测方法. 首先对销售量的影响因素进行了特征分析, 确定了训练数据的特征和维度. 然后采用基于 Adaboost 的随机森林销量预测方法对特征数据进行训练并给出了预测算法的步骤. 最后使用 python 进行了仿真实验, 实验结果表明, 该方法可以有效提高随机森林的回归性能, 且预测精度高, 具有较强的泛化能力.

关键词: Adaboost 算法; 随机森林; 销售量预测; 弱预测模型; python

引用格式: 常晓花,熊翱.基于 Adaboost 的随机森林算法在医疗销售预测中的应用.计算机系统应用,2018,27(2):202–206. http://www.c-s-a.org.cn/1003-3254/6203.html

Application of Random Forest Algorithm in Medical Sales Forecast Based on Adaboost

CHANG Xiao-Hua, XIONG Ao

(Institute of Network Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China)

Abstract: A sales forecasting method based on random forest algorithm and Adaboost method is proposed. Firstly, by analyzing the characteristics of the sales factors, the characteristics and dimensions of the training data are determined. Then, the feature data is trained by the random forest algorithm based on Adaboost, and the steps of the prediction algorithm are presented. Finally, the experimental results show that this method can greatly enhance the performance of random forest algorithm, and has a high prediction accuracy, as well as a good performance of generalization.

Key words: Adaboost algorithm; random forest algorithm; sales forcasting method; weak prediction model; python

1　引言

销售量是衡量一个公司销售业绩好坏的重要标志. 销售量的多少不仅影响着公司的盈利和发展, 同时也影响着整个国家的经济命脉. 因此, 销售量预测的精确性和科学性具有重要的研究价值.

销售数据是一种动态的、非线性的、不规则的时间序列数据, 受季节气候、突发事件、经销商的销售能力、下级经销商的数量等等各种因素的影响. 国内外学者对各行各业销售量预测的方法, 主要有神经网络预测法[1]、聚类预测[2]方法等. 传统的 BP 神经网络方法, 容易陷入局部极小值, 而达不到预测的精度; 基于聚类的销量预测方法, 初始聚类中心的选择对聚类结果的影响较大, 并且需要不断调整聚类中心, 所以数据量较大时, 算法时间开销也会非常大.

本文针对上述问题, 提出了基于改进 Adaboost 算法[3]的随机森林预测方法, 该方法不存在神经网络算法会陷入局部极小值的缺点, 且在数据量较大和较小时都能够维持一定的预测精度, 对不平衡数据集来说, 能够平衡误差, 同时训练速度快, 能够实现并行化. 本文利用该改进的 Adaboost 随机森林预测方法, 对某医疗器械销售公司的实际销售数据集进行了仿真实验, 实验结果证明了本文所提方法的有效性.

① 收稿时间: 2017-05-13; 修改时间: 2017-05-31; 采用时间: 2017-06-08

© 中国科学院软件研究所 http://www.c-s-a.org.cn

2 随机森林算法和 Adaboost 算法

2.1 随机森林算法原理

随机森林[4]是 Bagging 算法和改进的决策树算法的结合.

Bagging 算法是多个个体弱学习器各自学习, 然后通过集合策略来得到最终的强学习器, 个体弱学习器之间不存在依赖关系, 个体弱学习器的训练集通过随机采样得到, 随机采样表示每次从训练集中采集固定个数的样本, 但是采集后都将样本放回; 普通的决策树算法, 会在节点所有样本特征中选择一个最优的特征来做决策树的左右子树划分. 但是, 随机森林中所用的决策树算法, 则是通过随机选择节点上的一部分样本特征, 再在其中选择一个最优的特征来做决策树的左右子树划分.

随机森林算法, 就是将改进的决策树算法作为弱学习器, 然后使用 Bagging 算法对弱学习器进行集成学习而得到的. 该算法结合了 Bagging 算法的随机采样以及决策树算法的随机特征选择, 这两个“随机”性, 因而其泛化能力强, 不容易陷入过拟合. 同时, 由于 Bagging 算法中弱学习器之间相互独立, 随机森林中的决策树可以并行学习, 因而, 随机森林算法的时间效率高.

随机森林的训练过程如下:

(1) 给定训练集 S, 测试集 T, 特征维数 F. 确定参数: 使用到的 CART 的数量 t, 每棵树的深度 d, 每个节点使用到的特征数量 f. 终止条件: 节点上最少样本数 s, 节点上最少的信息增益 m.

对于第 1–t 棵树, i=1–t:

(2) 从 S 中有放回的抽取大小和 S 一样的训练集 $S(i)$, 作为根节点的样本, 从根节点开始训练.

(3) 如果当前节点上达到终止条件, 则设置当前节点为叶子节点, 如果是分类问题, 该叶子节点的预测输出为当前节点样本集合中数量最多的那一类 $c(j)$, 概率 p 为 $c(j)$ 占当前样本集的比例; 如果是回归问题, 预测输出为当前节点样本集各个样本值的平均值. 然后继续训练其他节点. 如果当前节点没有达到终止条件, 则从 F 维特征中无放回的随机选取 f 维特征. 利用这 f 维特征, 寻找分类效果最好的一维特征 k 及其阈值 th, 当前节点上样本第 k 维特征小于 th 的样本被划分到左节点, 其余的被划分到右节点. 继续训练其他节点.

(4) 重复 (2), (3), 直到所有节点都训练过了或者被标记为叶子节点.

(5) 重复 (2), (3), (4), 直到所有 CART 都被训练过.

利用随机森林的预测过程如下:

对于第 1–t 棵树, i=1–t:

(1) 从当前树的根节点开始, 根据当前节点的阈值 th, 判断是进入左节点还是进入右节点, 直到到达, 某个叶子节点, 并输出预测值.

(2) 重复执行 (1), 直到所有 t 棵树都输出了预测值. 如果是分类问题, 则输出为所有树中预测概率总和最大的那一个类, 即对每个 $c(j)$ 的 p 进行累计; 如果是回归问题, 则输出为所有树的输出的平均值.

随机森林算法包含在 python 的 sklearn 学习包中, sklearn 是一个 python 的科学计算库, 里面对一些常用的机器学习算法进行了封装. 随机森林在 sklearn 中的回归类是 RandomForestRegressor, 随机森林需要调整的参数包括两部分, 第一部分是 Bagging 框架的参数如表 1 所示, 第二部分是 CART 决策树的参数如表 2 所示.

表 1 Bagging 框架的参数

参数	含义
n_estimators	弱学习器的最大迭代次数. 默认是100.
oob_score	袋外样本用来评估模型的好坏. 默认False. 一般设置为True, 用于反应一个模型拟合后的泛化能力.
criterion	CART树做划分时对特征的评价标准. CART分类树默认是基尼系数gini, 另一个是信息增益. CART回归树默认是均方差mse, 另一个是绝对值差mae. 一般默认均方差.

表 2 CART 决策树的参数

参数	含义
max_features	RF划分时考虑的最大特征数, 默认是“None”.
max_depth	决策树最大深度, 一般不限制, 如果模型样本量多, 特征也多的情况下, 要限制最大深度, 常用取值为10-100之间.
min_samples_split	内部节点再划分所需最小样本数, 如果某节点的样本数少于min_samples_split, 则不会继续再尝试选择最优特征来进行划分. 默认是2.
min_samples_leaf	叶子节点最少样本数, 这个值限制了叶子节点最少的样本数, 如果某叶子节点数目小于样本数, 则会和兄弟节点一起被剪枝. 默认是1.
min_weight_fraction_leaf	叶子节点最小的样本权重和, 如果小于这个值, 则会和兄弟节点一起被剪枝. 默认是0,
max_leaf_nodes	最大叶子节点数, 通过限制最大叶子节点数, 可以防止过拟合, 默认是None,
min_impurity_split	节点划分最小不纯度, 默认值1e–7.

2.2 Adaboost 算法原理

Adaboost 是一种迭代算法. 其主要过程分为三步,

© 中国科学院软件研究所 http://www.c-s-a.org.cn

首先对训练样本进行权重初始化; 然后对弱学习器进行训练, 如果某个样本点的预测值达到了所要求的精度, 那么在构造下一个训练集时, 它的权重就降低, 相反, 则权重提高, 接着权重更新过的样本集被用于训练下一个学习器, 整个训练过程如此迭代进行; 最后将各个学习器组合成强学习器[5].

Adaboost 算法最终得到的强学习器, 由各个弱学习器与其权重结合而成, 弱学习器的权重由每个弱学习器的误差率决定, 误差率小的弱学习器, 其权重则大, 这使其在最终的强学习器函数中起较大的决定作用.

因此, Adaboost 算法中的个体学习器之间有很强的依赖关系, 也因此导致其有预测精度高的优点. 并且, Adaboost 算法[6]提供的是一个学习框架, 这使其应用非常灵活, 可以使用各种分类回归模型来构建弱学习器.

2.3 基于改进 Adaboost 的随机森林销售量预测模型

本文将随机森林算法与 Adaboost 算法框架结合, 提出了一种改进 Adaboost 算法的随机森林销售量预测算法, 以下称为预测算法.

算法首先对训练的数据进行了特征选取, 随后对 N 个训练样本做了权重的初始化, 初始化权重值为 $1/N$, 接着将训练样本用随机森林算法进行学习, 得到 N 个训练样本的预测值;由预测值计算得到 N 个训练样本的相对误差值, 进而通过与既定误差阈值的比较, 计算出该弱学习器的预测错误率ε_t, 以及该预测器所有样本的平均相对误差值γ_t; 根据 Adaboost 算法[7]的思想, 当样本预测正确时, 则要减小该样本在下一个弱学习器中的权重, 因而以$\beta_t = e^{\varepsilon_t^2} + k \times \frac{1}{1+e^{-\gamma_t}}$作为样本权重系数, 其中$k \times \frac{1}{1+e^{-\gamma_t}}$为影响因子, k 为影响系数; 当预测正确率越小, 也即是错误率就越大时, 则增大下一个弱预测器的权重, 以让下一个弱预测器对上一个弱预测器错误预测的样本更重视, 即ε_t(即错误率) 越大, γ_t(平均相对误差) 就越大, 最终β_t就越大; 最后将所有弱学习器的预测结果加权平均, 就得到了最终的强预测器, 然而, 由于最终要让预测值更为精确, 所以要减小错误率高的弱预测器的权重, 而增大错误率低的弱预测器权重, 因此以β_t的倒数作为弱预测器的权重.

该算法详细步骤如下:

(1) 样本数据选择及网络初始化. 根据训练数据, 选择训练样本的数据特征. 训练样本数据为 (x_1,y_1), $(x_2,y_2),\cdots,(x_n,y_n)$, 其中, $x_i \in X$, X 表示某个训练数据集或者实例空间, $y_i \in Y$, Y 表示实际输出值. 令迭代次数 $M=1$时, 样本权重初始化为$D_t(i)=\frac{1}{n}, i=1,2,3,\cdots,n$, 其中 n 为训练样本数据集的数量. 初始化分类器的预测率ε_t, 令$\varepsilon_t = 0$. 初始化随机森林的初始阈值$\phi(0<\phi<1)$.

(2) 随机森林弱预测模型[8]预测. 通过选取合适随机森林算法参数值, 构造不同的随机森林弱预测器, 对于$t=1,...,T$进行 T 次迭代, 在训练第 t 个弱预测器时, 使用随机森林回归类对数据进行训练, 建立回归模型 $g_t(x) \to y$. 具体步骤如下:

1) 计算样本的相对误差err_{t_i}, $err_{t_i} = \left|\frac{g_t(x_i)-y_i}{y_i}\right|$, 令 $\gamma_t = \frac{\sum|g_t(x_i)-y_i|}{\sum I}$, 表示弱预测器的样本平均误差;

2) 计算第 t 个弱预测器的预测错误率ε_t, $\varepsilon_t = \frac{n_{e_{t_i}>\phi}}{N}$, 表示第 t 个弱预测器预测结果中相对误差大于ϕ的样本数量占总样本数量的百分比, 令 $\beta_t = e^{\varepsilon_t^2} + k \times \frac{1}{1+e^{-\gamma_t}}$, k 为调整系数;

3) 计算第 t+1 个预测器的训练样本权重$D_{t+1}(i)$, $D_{t+1}(i) = \begin{cases} D_t(i) \times e^{\beta_t}, err_{t_i} < \phi \\ D_t(i) \end{cases}$, 并将下一个分类器样本权重使用公式$\frac{D_t(i)-\overline{D_t(i)}}{\mathrm{MAX}(\mathrm{D_t(i)})-\mathrm{MIN}(\mathrm{D_t(i)})}$进行归一化处理, 作为第 t+1 个分类器的样本权重值;

4) 计算第 t 个分类器权重ω_t, $\omega_t = \frac{\frac{1}{\beta_t}}{\sum_{t=1}^{T}\frac{1}{\beta_t}}$, 归一化权重;

5) 进行第 t+1 次迭代;

6) 如果迭代次数达到 T, 则继续 7), 否则返回 1), 直到达到最大迭代次数;

7) 输出强预测器函数$g(x) = \sum_{t=1}^{T}\omega_t g_t(x)$;

最终强预测器函数的输出公式如下:

$$g(x) = \sum_{t=1}^{T}\omega_t g_t(x) = \sum_{t=1}^{T}\frac{\frac{1}{\beta_t}}{\sum\frac{1}{\beta_t}} g_t(x) = \sum\frac{\frac{1}{e^{\varepsilon_t^2}+\frac{k}{1+e^{-\gamma_t}}}}{\sum\frac{1}{e^{\varepsilon_t^2}+\frac{k}{1+e^{-\gamma_t}}}} g_t(x)$$

本文提出的基于 Adaboost 的随机森林预测算法, 结合了随机森林算法和 Adaboost 算法的优点, 时间效率高, 泛化能力强, 同时由于 Adaboost 的迭代, 预测精度较高. 预测算法使用了弱学习器中每次学习后的预

© 中国科学院软件研究所 http://www.c-s-a.org.cn

测错误率以及样本的平均相对误差作为影响弱学习器权重的因子, 使得在下一个学习器中能够更注重上一个学习器预测误差较大的样本, 从而提高了预测的精度.

3 实验与结果分析

本文的实验数据来自某医疗器械销售公司的销售数据. 共有 240 组样本数据, 部分数据如表 3 所示, tmp_max 表示最高气温均值, tmp_min 表示最低气温均值, prob_rain 表示降雨概率 (已做处理), lstwk_sales 表示上周的销量, lsttwo_sales 表示上上周的销量, 单位为实际销售单位 EA.

表 3 实验部分数据

Id	tmp_max	tmp_min	prob_rain	lstwk_sales	lstwo_sales
1	18.5	10.535	40.5	37	151
2	16.9275	8.3575	38.64	103	37
3	16	9.9975	29.42	105	103
4	13.8225	5.8225	27.67	154	105
5	12.0425	3.75	33.57	84	154
6	9.4625	2.1425	37.11	144	84
7	9.18	0.895	30.24	86	144
8	8.8575	0.68	28.75	145	86
9	7.4275	0.325	20.79	248	145
10	7.07	0.4625	23.64	265	248

选取样本中的 150 组数据进行训练, 其余 90 组作为测试样本. 然后用本文提出的方法进行训练. 实验采用 python [9,10]进行仿真, 实验过程中, 随机森林决策树个数设为 100, 随机森林其余参数均选择默认值, 构成弱预测器. 根据本文提出的预测方法, 设置 Adaboost 算法的最大迭代次数 T 为 20, 调整系数 k 为 1.1, 相对误差率阈值设 0.1.

表 4 所示为部分预测值及相对误差, 图 1 是使用改进 Adaboost 方法和未使用 Adaboost 方法的预测结果对比图. 实验结果表明, 预测数据误差率小于 0.1, 使用基于 Adaboost 改进后的算法能够比改进之前的预测率提高约 12%, 证明了本文提出的算法的有效性.

表 4 室内光流扰动效应检测

改进前	改进前误差	改进后	改进后误差	实际值
179.2548	0.046517	185.3247	0.01423	188
160.9512	0.069585	147.191	0.021856	150.48
125.49	0.08181	119.9497	0.03405	116
178.57	0.065079	187.8495	0.016495	191
183.0796	0.041468	187.3638	0.019038	191
177.65	0.110313	165.6016	0.03501	160
166.8748	0.076612	164.6478	0.062244	155
133	0.081301	129.2764	0.051027	123
126.52	0.171481	115.0811	0.065565	108
111.08	0.008214	121.6306	0.085987	112
173.99	0.107744	185.204	0.050236	195
214.7492	0.036999	217.6441	0.024018	223
200.8248	0.107445	215.4068	0.042637	225
103.7648	0.080883	99.90034	0.040629	96
165.3	0.156633	188.134	0.040133	196
180.62	0.139905	195.1256	0.070831	210
129.87	0.139211	118.7265	0.041461	114
236.474	0.146303	230.9626	0.1662	277
157.17	0.218372	135.9075	0.053547	129
125.58	0.308125	113.8403	0.185837	96

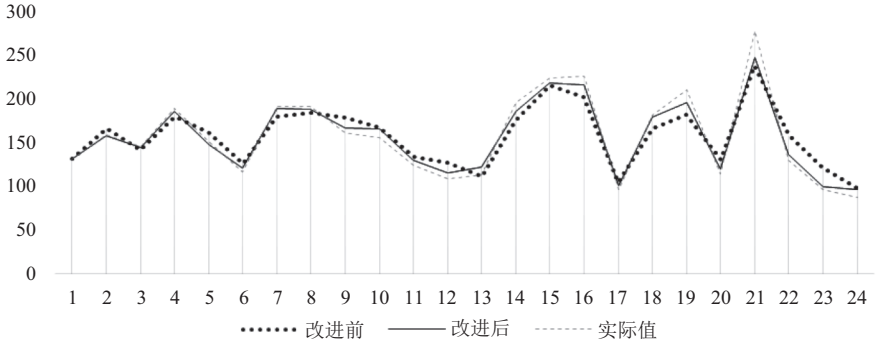

图 1 改进前后销量预测对比图

4 结论与展望

近年来, 随着医疗技术的不断更新发展, 医疗器材的需求量也不断增加, 大量的销售数据销售信息有待挖掘, 本文利用改进的 Adaboost 方法, 初始化训练样本权重, 并利用随机森林算法训练得到若预测器, 后又结合多个弱预测器, 形成强预测器的方法, 对销售数据集进行了分析和研究, 同时提升了随机森林算法的回归性能, 但是本文提出的方法, 只是通过改进 Adaboost 算法结合随机森林算法来提高算法的回归性能, 并没有考虑通过改进随机森林算法本身来提高算法的有效

© 中国科学院软件研究所 http://www.c-s-a.org.cn

性, 所以未来还有很大的提升改进空间.

参考文献

1 陈蓉. 基于 BP 神经网络的零售产品销量预测方法. 经营管理者, 2015, (4): 10–11.

2 王建伟. 基于商品聚类的电商销量预测. 计算机系统应用, 2016, 25(10): 162–168. [doi: 10.15888/j.cnki.csa.005423]

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

4 丁君美, 刘贵全, 李慧. 改进随机森林算法在电信业客户流失预测中的应用. 模式识别与人工智能, 2015, 28(11): 1041–1049.

5 李翔, 朱全银. Adaboost 算法改进 BP 神经网络预测研究. 计算机工程与科学, 2013, 35(8): 96–102.

6 张禹, 马驷良, 张忠波, 等. 基于 Adaboost 算法与神经网络的快速虹膜检测与定位算法. 吉林大学学报 (理学版), 2006, 44(2): 233–236.

7 李翔, 朱全银. 基于 Adaboost 算法和 BP 神经网络的税收预测. 计算机应用, 2012, 32(12): 3558–3560, 3568.

8 李威威, 李春青, 聂敬云, 等. 膜生物反应器膜污染的随机森林预测模型. 计算机应用, 2015, 35(SI): 135–137.

9 Richert W, Coelho LP. Python 语言构建机器学习系统. 南京: 东南大学出版社, 2016.

10 Hetland ML. Python 基础教程. 2 版. 司维, 曾军崴, 谭颖华, 译. 北京: 人民邮电出版社, 2014.

www.c-s-a.org.cn

© 中国科学院软件研究所 http://www.c-s-a.org.cn