文章编号：2095-2163(2020)02-0103-04 中图分类号：TP393 文献标志码：A

基于改进 AdaBoost 算法的选股模型

贺 超，吴 飞，何 洋，朱 海
（上海工程技术大学 电子电气工程学院，上海 201620）

摘 要：随着改革开放后国民经济的迅速增长，股票市场也日渐繁荣，因此量化投资交易技术也愈发受到重视。利用 AdaBoost 算法的选股模型虽然能够很好地达到预期效果，但是由于 AdaBoost 算法对异常值较为敏感，并且子分类器的决策结果权重对于最终结果有较大的影响，所以本文提出新的判决式特征选择机制以在训练阶段提高子分类器的鲁棒性，并且利用新的投票决策机制，结合了子分类器自身的精度和特征属性权重信息，使得算法整体结果得到提升。实验对比了 SVM 算法和传统的 AdaBoost 算法，结果表明所提出的改进 AdaBoost 选股模型有很好的效果。

关键词：选股模型；AdaBoost 算法；判决式特征选择；投票机制

Stock selection model based on improved AdaBoost algorithm

HE Chao, WU Fei, HE Yang, ZHU Hai
(School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai 201620, China)

【Abstract】 With the rapid growth of the national economy after the reform and opening up, the stock market is also increasingly prosperous, so quantitative investment trading technology has been increasingly valued. While using AdaBoost algorithm to pick stocks model can well achieve the desired effect, but since the AdaBoost algorithm is more sensitive to outliers, and the decision result weight of the component classifiers has great influence to the final result, so this paper puts forward the new sentence type feature selection mechanism for improving the robustness of the component classifiers in the stage of training, and uses the new voting judgment mechanism, combined with the feature of the classifier's own accuracy and attribute weight information, improves the overall algorithm results. The experimental results show that the improved AdaBoost stock selection model is effective.

【Key words】 stock selection model; AdaBoost algorithm; judgment selection; voting scheme

0 引 言

随着改革开放的不断深入，股票市场呈现出强劲崛起态势，并且在高速发展的当代中国社会扮演着重要角色。股票投资的主要目的就是在控制一定风险的前提下取得投资的最高收益。

传统的交易模式通常基于人为经验的对 MACD、BOLL 和 RSI 等技术指标进行判断，从而做出投资决策。由于大数据、云计算以及人工智能等科学技术的进步，传统的金融交易也深受影响，并且在实际量化投资领域运用中取得了良好效果。一直以来，股票市场吸引了各界的广泛关注与探讨研究，究其原因就在于其具有各种复杂多变的指标和观测角度，使得投资机遇与风险并存。支持向量机（Support Vector Machine，SVM）是基于统计学习理论推演生成的数据挖掘技术[1]，但是由于 SVM 对于大数量级的数据样本的训练有一定的难度，而实际面临的股市信息数据巨大，所以传统的 SVM 方法不足以支撑大规模训练强度。

针对股票信息受到影响波动拐点较多等特点[2]，单独的分类或预测算法无法做到较为灵活处理的问题，经过研究可知，AdaBoost 算法通过权重结合若干个弱分类器进行串行的学习[3]，并且通过联合权重投票机制求得最终结果。同时考虑到股票因子繁杂，受到较多因素影响，如此一来就会在样本数据集层面上引入较多的不确定性噪声，而 AdaBoost 算法对于异常值较为敏感，对于最终结果也会造成较大的影响[4]，所以在训练阶段选用了判决式的特征因子选择方法，能够在一定程度上剔除相关影响，与传统决策机制相比[5]，除了分类器自身的精度信息外，还充分利用了特征因子权重信息来辅助决策，使得整体效果得到了显著提升。利用上述分析来研究上市公司的财务指标与个股价格浮动率之间的关系，从而建立选股分类模型[6]。这里对此课题拟展开研究论述如下。

1 AdaBoost 算法

自适应增强算法（Adaptive Boosting

基金项目：国家自然科学基金(61272097)；上海市科技学术委员会重点项目(18511101600)。

作者简介：贺 超(1995-)，男，硕士研究生，主要研究方向：数据挖掘、最优化；吴 飞(1967-)，男，博士，教授，主要研究方向：计算机网络与计算机、能耗优化。

收稿日期：2019-12-25

Algorithm)[7],即 AdaBoost 算法,其主要思想是对于股票样本训练集合 $D=\{(x_1,y_1),(x_2,y_2),\dots,(x_N,y_N)\}$,其中 x_i 表示股票样本的因子属性特征,y_i 表示个股的输赢率作为标签变量,N 表示样本个数,以股票一年为时间节点的后复权股价涨跌幅大于 HS300 指数的涨跌幅取“1”,小于则取“0”,所以有 $Y\in\{+1,-1\}$。在选定好弱分类器后,初始状态下,所有样本权重相等,根据 AdaBoost 思想,不断串行迭代训练,并且在训练过程中后一个弱分类器将会着重训练被前一个弱分类器错分的样本,最终得到加权后的最终结果[8]。此处,给出主要流程具体如下。

输入:(x_1,y_1),$(x_2,y_2),\dots,(x_N,y_N)$,其中 $x_i\in X$,且 $y_i\in Y$

初始化:$\boldsymbol{W}^{<1>}=(w_1^{<1>},w_2^{<1>},\dots,w_N^{<1>})^{\mathrm{T}}$,$w_i^{<1>}=1/N$,其中 $i=1,2,\dots,N$,表示第 i 个分类器样本的权重分布。

训练过程:

for m in range (M):

Step 1 利用具有权重向量 $w_i^{<m>}$ 的训练数据集对弱分类器进行训练,其中 m 表示基分类器的个数,得到基分类器,可表示为公式(1):

$$h_m(X):x->\{-1,1\}, \tag{1}$$

Step 2 通过 $h_m(X)$ 在训练集上的效果,计算分类误差率,可表示为公式(2):

$$e_m=P(h_m(x_i)\neq y_i)=\sum_{i=1}^{N}w_i^m*I(h_m(x_i)\neq y_i), \tag{2}$$

并且,若分类误差率 $e_m\geqslant 1/2$,则算法提前停止,整体构建失败。

Step 3 为基分类器分配相应的构建权重系数,可表示为公式(3):

$$\alpha_m=\frac{1}{2}*\log\frac{1-e_m}{e_m}, \tag{3}$$

Step 4 更新训练权重向量 $\boldsymbol{W}^{<m+1>}=(w_1^{<m+1>},w_2^{<m+1>},\dots,w_N^{<m+1>})^{\mathrm{T}}$,其中 $w_i^{<m+1>}$ 的数学公式可表示为:

$$w_i^{<m+1>}=\frac{w_i^{<m>}}{Z_m}\exp(-\alpha_m y_i h_m(x_i)),\ i=1,2,\dots,N, \tag{4}$$

而 $Z_m=\sum_{i=1}^{N}w_i^m\exp(-\alpha_m y_i h_m(x_i))$ 为规范化系数。

输出:构建若干基分类器的线性组合,有 $f(x)=\sum_{m=1}^{M}\alpha_m h_m(x)$,根据 AdaBoost 核心思想,组合得到最终强分类器为公式(5):

$$H(x)=sign\left(\sum_{m=1}^{M}\alpha_m h_m(x)\right). \tag{5}$$

2 改进 AdaBoost 算法

2.1 判决式因子选取

根据随机子空间(Random Subspace Method,RSM)树结构采样方法[3],主要是从整个数据集中随机采样得到每个子树空间的子样本集,每次在建立子分类器的过程中,并不是采用整个数据集作为输入,当数据样本数量足够大时,通过实验表明,此种策略最终得到的分类结果精度要高于传统的 AdaBoost 算法。但是,上述随机采样在多次采样过程中,会出现某些样本被多次重复提取,而某些样本仅有少量的机会、甚至在建模阶段未被采用的情况,这就会导致基分类器的多样性受到制约。

受到前文两种现象的启发,采用判决式因子选择方式。假设给定的数据集 D 中有 N 个样本,如果在缺少特征 a_i 的状态下导致本轮基分类器的错误分类个数为 n,则认为本轮数据集属性判决 $J_i=n$。相反,如果因为缺少特征属性 a_i,错误分类个数为 n,本轮数据集属性判决为 $J_i=-n$,此种现象表明,某个属性的判决值的绝对值越大,则该属性对于整个数据集的重要程度越高。所以,为了提高子树之间的多样性,从所有特征属性中选择前 T 个属性作为数据集 D_k 用于创建子树 C_k,并且所有属性的权重初始化为 $\frac{1}{ac}$,其中 ac 表示对于训练集的特征属性个数,并且 T 由经验可启发式地设置为 $\frac{ac}{2}$。进一步地聚焦至单个节点 d 的属性划分选择,随机选取 ar($ar<T$)个特征属性用于计算其基尼系数,其中 ar 可表示为:

$$ar=1+\log_2 T, \tag{6}$$

研究中,并不是选择整个数据集的所有特征进行计算,选择基尼系数小的特征属性作为分割点,可表示为:

$$G[g(a_j(d))]=gini(d)-gini(a_j(d)),\ j\in[1,T], \tag{7}$$

其中,$gini(d)$ 表示该节点分割前的基尼系数,对应的 $gini(a_j(d))$ 表示在节点 d 中以最佳特征属性 a_j 分割后的基尼系数。

由于采取特征属性随机采样的机制，就使得在构建基分类器的过程中会出现某些特征属性被多次采取的情况，而在样本个数相同的前提条件下，从特征属性采样的角度来分析，就势必造成了数据的不均衡，因此当所在基分类器建成后，对于被多次选择的特征属性 a_j，可进行如下处理：

$$\mu(G[g(a_j(d))]) = \frac{G[g(a_j(d))]}{ns(a_j)}. \quad (8)$$

其中，$ns(a_j)$ 表示选择特征属性 a_j 的次数，$\mu(G[g(a_j(d))])$ 表示其均值，在子决策树中选择所有 $G[g(a_j(d))]$ 和其对应的 m 个特征属性（$m \leqslant T$），可推导计算出整体对应的均值 $\mu(G(g))$ 和标准差 $\sigma(G(g))$，并且如果 $\mu(G(g))$ 和 $\sigma(G(g))$ 之间的差值是正数，则提高特征属性 a_j 的权重，反之减少其对应的权重。

2.2　改进决策机制

由 2.1 节内容可知，为了保证子树之间的多样性，改进 AdaBoost 算法对于样本特征属性进行随机采样，并不是完整使用样本的所有数据，对子分类器进行训练，从而提高了各子分类器之间的多样性，更贴近真实数据多变的情况。

改进 AdaBoost 算法采用包外估计的方法，选用 2/3 的训练数据用于构建子树，即基分类器，此外 1/3的数据用于模型建成后的验证及相关学习权重的验证。利用训练数据集 D_k 去构建子树基分类器 C_k，将测试数据作为输入时，由前述切割原理可知，通过计算特征属性的基尼系数得到最佳切割属性 a_j，再将测试数据通过基分类器得到分类结果的平均精度作为子树基分类器 C_k 的属性 a_j 的决策权重 $w_{k,j}$。而在真正的在线使用阶段，对于任何一个未知的样本属性，改进后的算法将综合考虑属性分割点 a_j 的决策权重 $w_{k,j}$ 和子分类器的自身精度去计算最终的联合投票权重，最终分类预测结果可表示为：

$$I - AdaBoost(x) = \max_{\substack{i=0 \\ C_i(x)=y}}^{k} \sum \log\left(\frac{1}{1-(acc_i * w_{ij})^{\frac{1}{2}}}\right), \quad y \in Y. \quad (9)$$

其中，$I - AdaBoost(x)$ 表示改进算法的预测结果；y 表示真实的分类标签；$C_i(x)$ 表示子树基分类器的预测结果；acc_i 为子树 C_i 的精确度；w_{ij} 即为切割属性 a_j 的决策权重。

通过新的决策集成机制，充分保留了对特征属性随机采样而形成的子树之间的多样性，并且结合传统的投票决策方式，在提高预测结果精确度的同时，更好地切合了真实数据不确定性和多变性，从而有效提升了模型的鲁棒性。

3　实验设计与分析

3.1　实验设计

本文基于同花顺平台提供的 iFinD 数据库接口，以 HS300 为股票池，提取了 2008～2018 年的年度每只股票财务指标数据。文中例举了贵州茅台的财务指标实验数据见图 1。

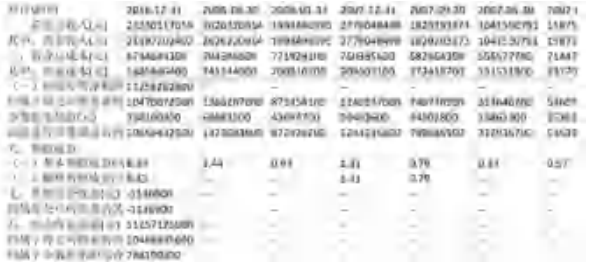

图 1　贵州茅台的财务指标实验数据

Fig. 1　Experimental data of financial indicators of Moutai, Guizhou

实验选取 2008～2018 年 HS300 为股票池中的股票数据作为实验数据，实验数据为每只个股的财务指标数据，包含营业总收入、营业总成本、营业利润、利润总额、净利润、每股收益、其他综合收益、综合收益总额等信息。目标函数是通过计算每只个股复权股价涨跌幅是否大于 HS300 指数涨跌幅计算求得。如果个股指数涨跌幅大于 HS300 指数的涨跌幅则取“1”，小于则取“0”，实验以 2008～2017 年数据为训练数据集，以 2018 年数据作为测试数据集。

3.1.1　评价标准

对于改进 AdaBoost 模型，在实际运用中，以分类准确率为其性能好坏的评价标准，其数学定义可写为：

$$p = \frac{\text{测试数据集分类正确的数目}}{\text{测试数据集样本总数}}. \quad (10)$$

3.1.2　设计流程

股票投资中，股票收益率的涨跌幅是一个非常重要的指标。根据模型规则，如果预测下一年的收益率为正，则做出买入的决策，并且投资状态设置为 1；如果预测下一年的收益率为负，则做出卖出的决策，并且投资状态设置为 0。决策流程如图 2 所示。

3.2　实验分析

在量化交易发展初期，SVM 算法由于其原理的简单易用性，在实际运用中取得了很好的效果，但是随着数据量级的增加，SVM 在大数量级的交易数据和研报数据的处理中暴露出不足之处，这也是其算

法本身存在的问题。由于 AdaBoost 算法框架思想的提出,使得可以集中各弱分类器,并在每一步中不断地进行迭代优化,因为其对异常值较为敏感的因素,在实际生产数据的应用上会产生较大的影响,因此对于传统的 AdaBoost 算法,加入新的特征属性选择机制,如此即使得最终的决策机制同时结合了子分类器自身的精度和特征属性权重信息,使得最终的分类精准度得到了极大的提升。本次研究中各选用算法的结果对比曲线如图 3 所示。

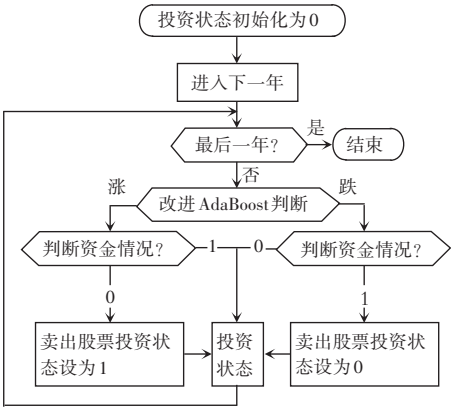

图 2 决策流程图

Fig. 2 Decision flow chart

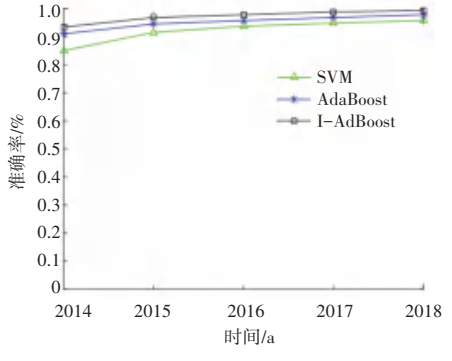

图 3 分类准确率

Fig. 3 Classification accuracy

由图 3 分析指出,由于改进后的 AdaBoost 算法融合了属性自身精度和基分类器的精度,更加贴合实际决策方式,提高了系统的鲁棒性,而相比于传统的 AdaBoost 算法,SVM 性能上要稍有逊色。改进后的 AdaBoost 算法的实测效果最佳,分类准确度可达到99.3%。

上述对比主要是基于业务层面的分析,下一步则需讨论模型本身的性能分析,而为了更好地分析 3 种算法模型的性能,选取 2014~2018 年间的数据作为样本,分析对比结果如图 4 所示。

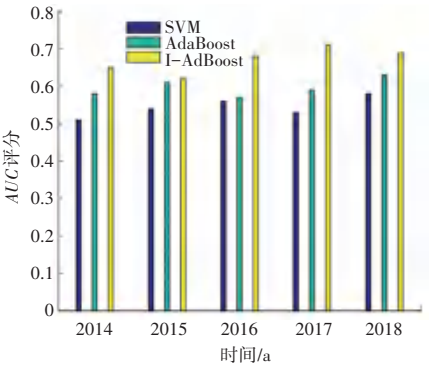

图 4 *AUC* 评分

Fig. 4 *AUC* score

由图 4 分析可知,从每个时期上看,因为改进后的 AdaBoost 算法运用新的判决式因子选择机制,保证了基分类器间的多样性,提高了算法整体的鲁棒性,所以每个时期的 *AUC* 评分非常稳定,并且评分较高,最高评分可达 0.71,这就表明改进后的 AdaBoost 算法自身性能上较为稳定且有好的实际效果。其中,SVM 算法与传统的 AdaBoost 算法相比,性能上仍有欠缺。

4 结束语

随着中国一带一路等政策的发展,逐渐走向国际市场,股票市场将不断完善。金融科技的布局,也将给股票市场带来新的活力。本文从股票的投资价值角度分析,利用改进 AdaBoost 算法,通过新的判决式属性选择机制保持了基分类器的多样性,更客观地贴合实际股票数据的情况,增强了整体的鲁棒性,与此同时,在最终的投票机制中融合了特征因子自身的精确度和基分类器的精确度评分,很大程度上提高了最终的决策性能,在实际应用中有着良好的适用性。

参考文献

[1] VAPNIK V. The nature of statistical learning theory[M]. New York: Spring-Verlag, 1995.
[2] 黄秀霞, 孙力. 基于属性依赖度计算和 PCA 的 C4.5 算法[J]. 传感器与微系统, 2017, 36(1):131.
[3] 袁泉. Adaboost 组合分类模型在信用评估领域应用研究[D]. 哈尔滨:哈尔滨工业大学, 2011.
[4] 李云飞, 龚冬生, 惠晓峰. 基于价值投资的 PCA-SVM 股票选择模型研究[J]. 西安工程大学学报, 2009, 23(3):125.
[5] 李想. 基于 XGBoost 算法的多因子量化选股方案策划[D]. 上海:上海师范大学, 2017.
[6] 汪东. 基于支持向量机的选时和选股研究[D]. 上海:上海交通大学, 2007.
[7] 陈中杰, 蒋刚, 蔡勇. 基于 SVM 一对一多分类算法的二次细分法研究[J]. 传感器与微系统, 2013, 32(4):44.
[8] 全林, 姜秀珍, 赵俊和, 等. 基于 SVM 分类算法的选股研究[J]. 上海交通大学学报, 2009,43(9):1412.