文章编号:1007－6735(2018)03－0267－07　DOI: 10.13255/j.cnki.jusst.2018.03.009

基于随机森林的 A 股股票涨跌预测研究

林娜娜[1]，　秦江涛[2]
（1. 上海大学 管理学院，上海 200444；2. 上海理工大学 管理学院，上海 200093）

摘要：针对传统预测模型易陷入过拟合、缺失数据敏感、计算量大等不足，利用随机森林算法的双重随机性、处理数据集优异等特点，对 A 股股票涨跌预测进行研究。首先运用相关性分析对初始指标体系进行一次 Spearman 和二次 Pearson 筛选，去除指标体系中的冗余指标。然后对随机森林的各项重要参数进行优化，并对优化后的模型采用重要性估计方法以提升训练模型精确度。通过不同指标体系的对比，验证实验过程的正确性。最后，对比不同建模方法的实证预测结果，表明随机森林模型比传统机器学习方法二元 logistic 回归在性能上更优越，具备较高的预测准确度。

关键词：随机森林；股票；预测
中图分类号：TP 301.6　　**文献标志码：**A

Forecast of A–Share Stock Change Based on Random Forest

LIN Nana[1],　QIN Jiangtao[2]
(1. *School of Management, Shanghai University, Shanghai 200444, China*;
2. *Business School, University of Shanghai for Science and Technology, Shanghai 200093, China*)

Abstract: In view of the shortcomings of traditional forecasting models such as easy overfitting, missing data sensitivity and large computation, a random forest algorithm was used to study the stock price change forecast, utilizing its benefits of double randomness excellent data processing performances. First, the correlation index analysis was carried out to select the initial index system once and twice. Next, the important parameters of the random forest were optimized, and then an importance estimation method was adopted to improve the accuracy of the training model. Through the comparison between different index systems, the correctness of the experimental process was verified. Finally, comparing the empirical results of different modeling methods, it is shown that the random forest model is superior to the binary logistic regression model and has higher prediction accuracy.

Keywords: *random forest*; *stock*; *forecast*

收稿日期: 2017−07−02
基金项目: 国家自然科学基金资助项目(71401107)
第一作者: 林娜娜(1995–)，女，硕士研究生．研究方向：金融数据挖掘．E-mail：kally5_lin@163.com
通信作者: 秦江涛(1966–)，男，副教授．研究方向：信息管理、先进制造系统等．E-mail：qinjiangtao_usst@126.com

股票市场研究中，股票涨跌的预测一直是关注的热点。吴微等[1]通过BP神经网络算法对沪市综合指数涨跌情况进行预测，达到了良好的预测精度，但神经网络模型对个股的走势预测效果欠佳。刘道文等[2]采用基于支持向量机的股票选择模型，并以交叉验证法确定了最佳回归参数，并以此建立了预测模型，对上海证券交易所股票价格指数预测效果比较理想，但对于核函数和最佳参数的选取还有提升的空间。戴钟仪[3]运用关联规则，对沪深300指数成分股进行涨跌预测，论证了股票涨跌过程中存在着一定的规律。

经典机器学习算法之一的神经网络算法虽然预测较为精准，但是计算量繁琐。而支持向量机对缺失数据敏感，会极大地影响输出结果，不能适应目前股票预测模型的实际需要。因此，诸多学者建议采用模型组合的方式来提升预测准确度。随机森林算法是一种模型组合，应用到不同的领域上均获得不俗的成果[4-7]。基于随机森林算法的优势，将该算法运用到股票涨跌预测中，能够避免上述预测模型的不足。根据现有文献可知[8-10]，随机森林法预测主要是先对建立的初始指标体系进行筛选，将筛选后的指标数据作为影响变量代入到随机森林中，涨跌情况作为响应变量输出。但现有方法对随机森林本身的模型优化有所欠缺，不能进一步提升预测精确度。

本文在此基础上对随机森林算法进行系统性优化，通过对随机森林中的各项重要参数进行逐步测试，如树节点的变量数(简称：mtry)、树的个数(简称：ntree)、OOB(out of bag)误分率以及变量重要性估计等来提升预测准确度，从而得到预测模型，研究其对股票市场投资决策存在的实际应用价值。

1 指标体系构建

为了建立A股股票涨跌预测模型，首先要确定必要的影响指标作为模型的输入，必要的响应变量作为模型的输出，因为构建股票指标体系是进行后续评价和综合分析的基础。

1.1 初始指标体系建立

吴微等[1]选取成交额、成交量、涨跌幅等股票市场因子作为神经网络研究方法中的指标；国琳[11]等利用4个方面财务因子包括盈利能力、偿债能力、资产营运能力、成长能力运用到股票价格预测中，用实证分析说明其研究的实际价值；谢国强[12]选取股票的开盘价、最高价、最低价、收盘价等市场因子作为支持向量回归机的输入向量，证明该预测模型具有较好的预测精度和泛化能力。但无论是运用股票市场因子或者财务因子作为指标都较不全面，仍有需要改进的地方。

股票的涨跌问题是由复杂因素和环境共同影响导致的，所以本文的初始选股指标体系结合股票市场因子和财务因子共同作为初始指标体系。基于晓雯等[13]和许华丰等[14]对股票指标体系的设计原则，参考其他学者的指标选取，同时剔除了有缺失数据的指标之后，建立以下基本指标体系的初步框架，如表1所示。

表1 初始股票指标体系

Tab.1 Initial stock index system

因子类别	指标类别	指标名	简称
股票市场因子	价格指标	前收盘价	PC
		开盘价	OPEN
		最高价	PP
		最低价	BP
		收盘价	CLOSE
		均价	AP
	流通量指标	成交量	SV
		成交额	ST
		流通市值	MV
	投资效益	市现率	PCF
		市盈率	PE
		市销率	PS
		市净率	PB
	活跃程度	换手率	TR
		振幅	SA
	综合	复权因子	RF
		股权价值	SVA
		企业倍数	EV
		当月涨跌数	MF
		当月涨跌幅	MFR
财务因子	收益能力	基本每股收益	EPS
		净资产收益率	ROE
	偿债能力	流动比率	CR
	盈利能力	净利润率	NPR
		总资产报酬率	ROA
	经营能力	总资产周转率	TAT

1.2 影响变量与响应变量

以表1中26个指标作为初始影响变量，根据预测模型的未来性，将每只股票下月的涨跌数NMF与0比较，建立响应变量。当$NMF_i \geqslant 0$，归为一类，当$NMF_i<0$归为另一类。其中$i=1,2,\cdots,n$，NMF_i是第i只股票下月的涨跌数，n为选取行业的股票数。

2　相关性实证分析

进行相关性分析的 100 只样本股票，它们的 26 个影响变量选取时间为 2016 年 6 月和 2016 年 7 月的月线数据，响应变量选取时间对应为 2016 年 7 月和 2016 年 8 月月度涨跌情况，其中市场因子和财务因子的数据均来自于东方财富 Choice 金融客户端和中国证监会指定信息披露的网站：巨潮资讯网。这 100 只股票皆来自于软件和信息技术服务业，是信息传输、软件和信息技术服务业中的子行业。

2.1　一次筛选——Spearman 相关研究

为了通过筛选指标来建立合适的指标体系，达到最终对股票涨跌预测。采用 SPSS 数理统计软件进行 Spearman 相关分析，分析每个影响变量与响应变量之间的相关系数，显著性检验选择的是双尾检验。根据得出的相关系数数值大小，从大到小进行排序，如表 2 所示。

表 2　Spearman 相关分析结果

Tab.2　Spearman correlation analysis results

指标	相关系数	显著度	置信区间
当月涨跌幅	0.486	0.000	0.01
当月涨跌数	0.417	0.000	0.01
总资产周转率	(0.242)	0.001	0.01
换手率	0.204	0.004	0.01
成交额	0.192	0.006	0.01
总资产报酬率	(0.190)	0.007	0.01
收盘价	0.188	0.008	0.01
净资产收益率	(0.184)	0.009	0.01
成交量	0.167	0.018	0.05
市净率	0.163	0.021	0.05
基本每股收益	(0.158)	0.026	0.05
振幅	0.150	0.034	0.05
市销率	0.147	0.037	0.05
股权价值	0.143	0.043	0.05
最高价	0.134	0.058	
流通市值	0.129	0.070	
市盈率	0.128	0.072	
均价	0.123	0.084	
企业倍数	0.116	0.103	
最低价	0.099	0.162	
流动比率	0.097	0.172	
开盘价	0.091	0.200	
前收盘价	0.086	0.226	
净利润率	(0.082)	0.249	
市现率	0.046	0.517	
复权因子	(0.010)	0.883	

普遍情况下当显著度数值小于 0.05，表明其相关系数数值是可信的，而不是因为样本抽样误差所产生的。通过观察可以发现，显著度小于 0.05 的前 14 个指标的相关系数明显比后 12 个指标来得大。综合考虑各个指标的相关系数大小和对应的显著度，保留前 14 个影响指标加入到一次筛选后的指标体系中。

2.2　二次筛选——Pearson 相关研究

为了保证指标与指标间存在较高的相异性，去除指标间存在显著相关的冗余指标，得到精简的指标体系，从而更进一步优化模型。同样通过 SPSS 进行 Pearson 相关分析，分析一次筛选后的 14 个指标之间的相关系数。显著性检验选择的是双尾检验。去除标准：当两个影响变量之间高度相关时，考量这两个影响变量与响应变量之间的相关系数，舍去相关系数小的，保留相关系数大的。经过二次筛选后的新指标构建了本文的二次筛选 13 指标体系，如表 3 所示。

表 3　二次筛选股票指标体系

Tab.3　Secondary screening of stock index system

指标名	指标类别
收盘价	价格指标
成交量	流通量指标
成交额	
市销率	投资效益
市净率	
换手率	活跃程度
振幅	
股权价值	综合
当月涨跌数	
当月涨跌幅	
基本每股收益	收益能力
总资产报酬率	盈利能力
总资产周转率	经营能力

3　随机森林实证研究

随机森林算法在样本数据分布不平均、高纬度、存在部分特征缺失的情况下，仍能维持一定的准确度，并且可以在运算量没有明显增加的条件下提升准确度。在处理样本数据时同样表现出

较大优势，具备很好的抵抗噪音的能力以及不易陷入过拟合。随机森林参数中的 *mtry* 和 *ntree*，可用于协调分类精确性与多样性之间的平衡，OOB 误分率可用于对随机森林的泛化误差进行无偏估计，变量重要性估计能够计算每个特征变量对分类结果的重要性。因此，本文在实证研究中对随机森林中的各项重要参数进行测试，以提升预测准确度。

具体研究应用 ChiMerge 算法对原始数据进行预处理。ChiMerge 依赖于卡方分析，如果两个相邻区间的卡方值很低，则表明两者具有非常类似的类分布，那么这两个区间便能够合并；否则，它们应当保持独立，从而达到精确离散化的目的。

3.1 确定随机森林中 *mtry* 的值

二次筛选后的指标体系共有 13 个指标，分别实验得出当 *mtry* 为 1～13 时每次实验的误分率值。为了保证实验结果的有效性，重复以上实验多次，取每个 *mtry* 值的误分率均值作为判别 *mtry* 合适值的标准。实验结果如图 1 所示。

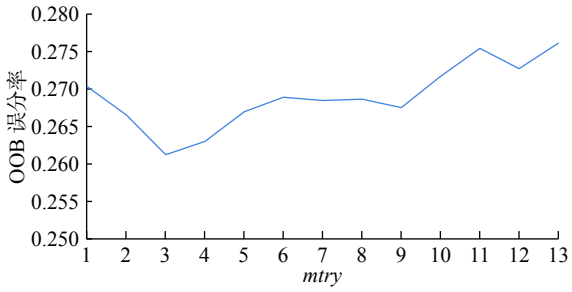

图 1 OOB 误分率均值变化

Fig.1 Change of the mean error rate

观察图 1 可知，当 *mtry*=3 时，误分率达到最低。符合目前的研究[15]，*mtry* 多取为$\sqrt{M}$，*M* 为指标总个数。

3.2 确定随机森林中 *ntree* 的个数

基于上述确定的 *mtry*=3，对随机森林另一个重要参数 *ntree* 进行随机建模，寻求一个适当的 *ntree*。通过实验，图形化展示误分率与树的数量之间的变化关系，如图 2 所示。

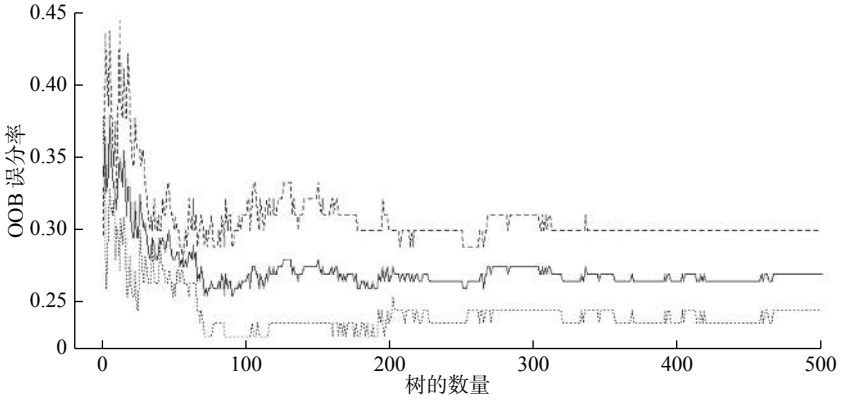

图 2 误分率与树的数量之间变化关系

Fig.2 Relationship between the error rate and number of trees

观察图 2 可知，当 *ntree*=200 之后，误分率趋于平稳且可以使模型分类精度达到要求。符合目前学者关于随机森林的研究[15]，*ntree* 值皆以大于 100 棵为合适。

3.3 划分区间数量的测试

区间数量对预测的准确度会有显著的影响。根据实验经验可得，本文的样本数量为 200，合适的区间数量为 20～29，运用 R 语言对应当在这个范围内划分多少区间数才能使 OOB 误分率最低进行研究。在基于 *mtry*=3 和 *ntree*=200，再调用经过 ChiMerge 离散化的 2016 年 6 月、7 月、8 月二次筛选后的 13 个指标数据，进行随机森林的 OOB 误分率测试。详细步骤如下：

a. 使用 ChiMerge 算法对连续性原始数据进行 20～29 个区间的离散化处理，以下步骤以 25 区间为例，其他 9 个区间同理可得。

在 R 语言中，加载 randomForest 包，编写语句训练一个随机森林，输出模型的 OOB 误分率。由于随机森林的随机性，为了保证测试结果的稳定与可靠，同一个实验重复 50 次，记录每一次的 OOB 值，如图 3 所示。由结果可知，划分 25 个区间时平均 OOB 误分率的值是 26.1%，训练模型的准确率是 73.9%。

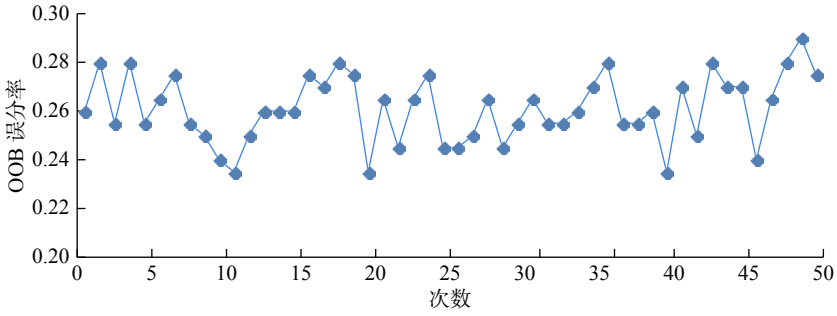

图 3　每个实验的误分率数值

Fig.3　Error rate of each experiment

b. 对上述过程做剩余 9 个区间实验，得出结果如表 4 所示。

表 4　OOB 误分率和精确度一览表

Tab.4　Model error rate and accuracy list

区间离散数	OOB 误分率均值	训练模型精确度均值
20 区间实验	0.263 2	0.736 8
21 区间实验	0.262 2	0.737 8
22 区间实验	0.256 7	0.743 3
23 区间实验	0.260 4	0.739 6
24 区间实验	0.260 8	0.739 2
25 区间实验	0.261 0	0.739 0
26 区间实验	0.262 0	0.738 0
27 区间实验	0.264 3	0.735 7
28 区间实验	0.264 9	0.735 1
29 区间实验	0.264 9	0.735 1

分析图 4 可以发现，随着区间数量的变化，随机森林模型的 OOB 误分率和精确度都在不断变化。随着区间数量的增加，模型的性能有所上升，达到 22 区间数量时性能最优，精确度达到 0.743 3，接着区间数量再增加之后，性能逐渐有所下降。所以，本文选定进行 22 区间数量划分时会使得随机森林模型的精确度达到最高。

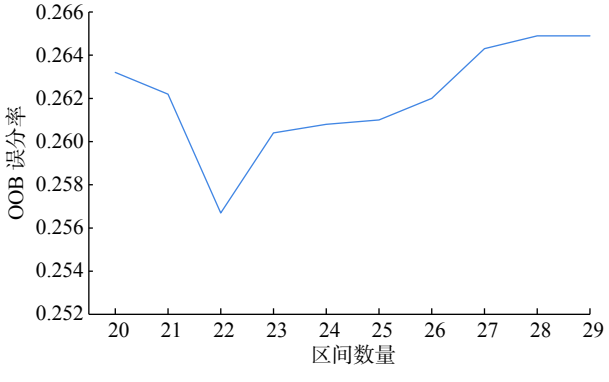

图 4　每个离散区间的误分率均值

Fig.4　Mean error rate in each discrete interval

4　重要性估计研究

4.1　重要性测试

在最佳 22 区间数量划分的基础上，对随机森林进行重要性估计测试。本文选用的重要性估计方法是基于 Gini 分类节点纯度下降量的方法。使用 R 语言 importance( )命令对模型指标进行重要性测试，得到指标重要性结果，如图 5 所示。

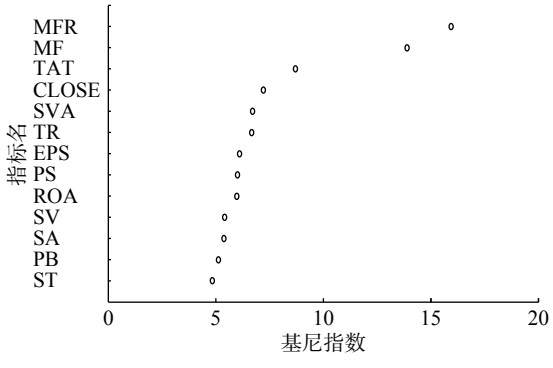

图 5　指标的重要性排序

Fig.5　Rank importance of indexes

4.2　重要性排序筛选

因为模型的特征变量中不乏噪音指标，这会影响随机森林算法的准确性，所以对每个特征的重要性值进行排序后逐步从重要性最低的开始剔除。目前股票预测中，最低指标数普遍为 6 个，过少的指标数量会导致模型的可解释性降低和预测结果偶然性的增加。

因此，本文在第一次实验中，筛除重要性最低的 ST 指标，第二次实验在前一次实验的基础上再筛除 PB 指标，第三次实验在前两次的基础上再筛除 SA 指标，依此类推，直到剩下 6 个指标为止，共实验 7 次。每个模型为了保证结果的稳定性重复 50 次，观察实验结果，如下页表 5 所示。

表 5 重要性排序筛选结果

Tab.5 Sort results by importance

实验名称	OOB 误分率均值	训练模型精确度均值
筛除 ST 指标	0.248 0	0.752 0
再筛除 PB 指标	0.249 5	0.750 5
再筛除 SA 指标	0.249 8	0.750 2
再筛除 SV 指标	0.233 8	0.766 2
再筛除 ROA 指标	0.223 3	0.776 7
再筛除 PS 指标	0.221 6	0.778 4
再筛除 EPS 指标	0.238 6	0.761 4

观察表 5 可知，进行重要性筛选能使训练模型精确度均值逐渐提高，但到达最大值后又开始减小。所以实验再筛除 PS 指标的训练模型精确度为 0.778 4，是 7 次实验中最优值。筛选后的指标体系包含 MFR，MF，TAT，CLOSE，SVA，TR 和 EPS 这 7 个指标，命名为重要性筛选 7 指标模型。

未经重要性排序筛选之前的最佳精确度均值只有 0.743 3，而经过重要排序筛选后的精确度均值提高了 3.51%。并且在 7 次实验中，每一次实验的模型精确度均值皆不同程度高于未经重要性排序筛选之前的精确度均值。由此可见，对随机森林进行重要性估计测试可提升模型精确度。

4.3 不同指标模型的 OOB 误分率对比

为了使对比结果便于观察，基于研究得出的最佳划分区间 22 个区间数为基础，对比 4 个指标模型的 50 次实验结果，可以得出结果，如图 6 所示。

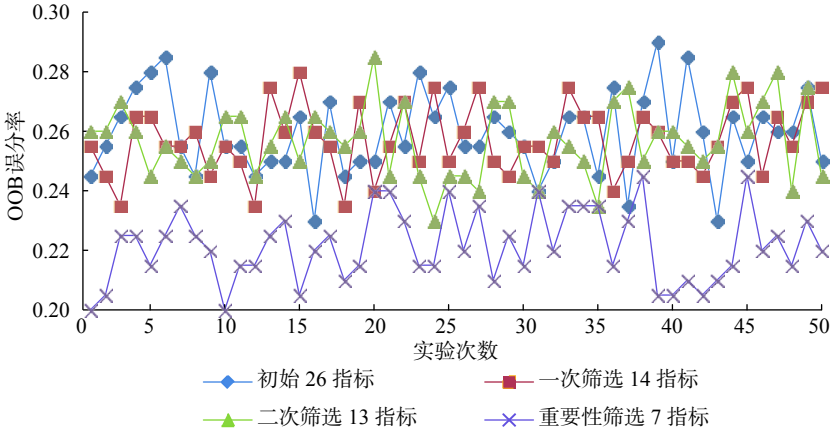

图 6 4 个指标模型的 OOB 误分率对比

Fig.6 Error rate comparison between four index models

使用 SPSS 统计软件对 4 个模型的 OOB 误分率数据进行描述性统计，结果如表 6 所示.

表 6 4 个模型的描述性统计

Tab.6 Descriptive statistics of four models

指标模型	实验次数	最小值/%	最大值/%	平均值/%	标准偏差/%
初始 26 指标	50	23.00	29.00	25.94	1.41
一次筛选 14 指标	50	23.50	28.00	25.72	1.17
二次筛选 13 指标	50	23.00	28.50	25.67	1.24
重要性筛选 7 指标	50	20.00	24.50	22.16	1.20

分析图 6 和表 6，可以得出：从模型的 OOB 误分率来看，一次筛选得出的 14 个指标整体比初始 26 个指标的 OOB 误分率低且更加稳定，表明随机森林在处理高纬度数据方面的优势，可以在一定程度上无需作特征选择；二次筛选的 13 个指标几乎包含了一次筛选后的 14 个指标带有的信息量，结果既相近又稳定；重要性筛选 7 指标模型的 OOB 误分率比上述 3 个模型的均低得多，且误分率较稳定。

由此可见，相关性研究大幅度地筛选出了冗余的指标，小幅度地提升了随机森林模型的精确度，在此基础上的重要性估计测试能够在很大程度上提高随机森林模型整体的精确度。

5 股票涨跌预测结果

5.1 随机森林涨跌预测结果

选定重要性筛选 7 指标模型作为本文的股票涨跌预测模型，将 2016 年 6 月与 7 月的影响变量离散化数据和对应 7 月和 8 月的涨跌情况代入到

模型中，对随机森林模型进行训练。根据训练之后的模型，将 8 月所需预测的数据代入，使用 randomForest 包中的 predict( )命令预测出 9 月股票市场的涨跌情况。

由表 7 进行统计分析可知，在样本 100 只股票中，对下月涨跌情况预测的总体精确度达到 83%，平均总体涨跌幅为–3.234 17。其中，预测将要上涨的 30 只股票，预测正确 22 只，预测精确度达到 73.33%，其涨跌幅均值为 1.578 04，比总体平均高了 4.812 21。预测将会下跌的 70 只股票，预测正确 61 只，预测精确度达到 87.14%，其涨跌幅均值为–5.296 55，比总体平均低了 2.062 38。从整体来看，该模型的性能优越，对协助投资决策有着实际应用价值。

表 7　样本股票的预测结果

Tab.7　Forecast results for the sample stock

名称	预测上涨的股票	预测下跌的股票	总体样本股票
股票数量	30	70	100
预测正确	22	61	83
涨跌幅最大值	29.758 94	8.943 09	29.758 94
涨跌幅最小值	–14.285 70	–20.992 40	–20.992 40
涨跌幅平均值	1.578 04	–5.296 55	–3.234 17

5.2　不同建模方法比较

为了进一步评价随机森林算法建立的模型，本文运用经典机器学习算法之一的二元 logistic 回归，通过相同的训练样本建立了回归模型，从而来比较不同建模方法所构建模型的预测能力。

通过研究图 7 可知，随机森林和二元 logistic 回归对训练模型的预测准确率水平相差不多，随机森林略高于二元 logistic 回归。但对测试样本进行预测方面，二元 logistic 回归预测准确度和稳定性明显不如随机森林。随机森林在数据集上表现出很大优势，因为取样随机和特性选择随机的引入，使得随机森林具备很好的抵抗噪音的能力，而且不易陷入过拟合。而二元 logistic 回归陷入了过拟合，虽然训练数据上能够较好拟合这些数据，但不能对于训练数据以外的其他数据进行很好的预测。

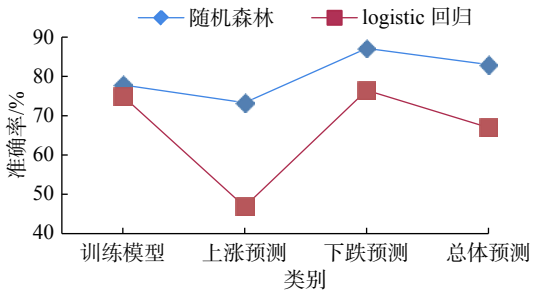

图 7　两种建模方法的准确率对比

Fig.7　Accuracy comparison of two modeling methods

6　结　论

本文采用机器学习中的随机森林算法以及 R 语言、SPSS 研究工具，对软件和信息技术服务业的 100 只股票 2016 年 9 月的涨跌情况进行预测。

主要通过相关性研究中的 Spearman 和 Pearson 方法，对初始指标体系进行筛选，剔除了冗余指标，精简了指标体系。运用 ChiMerge 算法对指标数据进行离散化，并基于随机森林的重要性估计方法逐次剔除重要性低的指标。最后，通过随机森林和二元 logistic 回归实证对比证明，随机森林算法的性能稳定且优越。本文建立的随机森林模型在软件和信息技术服务业取得了不错的预测结果，在接下的研究中，作者将继续探究随机森林在其他行业中的普适性，从而进一步论证实验结果。

参考文献：

[1] 吴微, 陈维强, 刘波. 用 BP 神经网络预测股票市场涨跌[J]. 大连理工大学报, 2001, 41(1): 9–15.
[2] 刘道文, 樊明智. 基于支持向量机股票价格指数建模及预测[J]. 统计与决策, 2013(2): 76–78.
[3] 戴钟仪. 关联规则在股票分析及预测中的应用[J]. 新经济, 2016(5): 59.
[4] 明均仁, 肖凯. 基于 R 语言的面向需求预测的随机森林方法[J]. 统计与决策, 2012(9): 81–83.
[5] 邓红卫, 陈超群, 张亚南. 岩体可爆性等级判别的随机森林模型及 R 实现[J]. 世界科技研究与发展, 2016(5): 946–949.
[6] 张雷, 王琳琳, 张旭东, 等. 随机森林算法基本思想及其在生态学中的应用——以云南松分布模拟为例[J]. 生态学报, 2014, 34(3): 650–659.
[7] CHAN J C W, PAELINCKX D. Evaluation of random forest and adaboost tree-based ensemble classification and spectral band selection for ecotope mapping using airborne hyperspectral imagery[J]. Remote Sensing of Environment, 2008, 112(6): 2999–3011.

（下转第 301 页）