基于机器学习算法的金融期权波动率预测*

马天平　吴卫星

内容提要　期权波动率预测是期权风险预警管理的关键问题，传统方法采取 GARCH 等时间序列模型。与传统方法不同，本文创建了基于机器学习算法的“SKRG 递进集成”新预警体系，体系以中国波指为对象，采取 48 个相关指标作为对中国波指预测的特征(Feature) ，依次引入 SVM 机器学习、KNN 样本不平衡机器学习、RF 划分、GBDT 优化完成机器学习建模过程，逐步提高预测精准率。测试样本显示，基于机器学习的预测效果好于传统的 GARCH 模型。本文的理论价值在于丰富了期权随机波动率预测领域的相关文献，应用价值在于为波动率的预测进而期权风险预警提供了新的方法。

关 键 词　机器学习　期权交易　波动率预测

DOI:10.16091/j.cnki.cn32-1308/c.2018.05.029

引　言

金融工程中，期权是重要的衍生品工具。作为机构交易者，在设计交易期权的策略中，突出的交易策略是卖出类。但单向卖出期权与单项买入期权一样，存在巨大的交易风险。为获取稳健的卖出类期权策略收益，需要动态对冲。

如何考虑对冲的动态连续性和前瞻性，成为风险管理的焦点。市场波动率是决定期权价格的重要变量，然而事实和研究表明，期权波动率并不是一成不变的，而是具有随机性。波动率的不可预测性意味着难以找到合适的波动率对期权予以定价。因而要把握期权价格的变化趋势以及对冲的动态性和前瞻性，对波动率的预测就成为十分重要的工作。比如，在卖出期权的策略中风险的很大一部分来自隐含波动率的大幅度上涨，因此如果我们能够提前预测出隐含波动率的上涨，便可以通过对冲仓位的调整来削减或是规避掉波动率上涨带来的风险。

波动率预测急需使用新的方法体系模型。近年来，随着大数据、人工智能、机器学习技术的日趋成熟，可以利用新技术实现波动率的预测。大数据是新技术处理模式中，具有更强的决策力、洞察力和流程优化能力的海量、高增长率和多样化的信息资产 T + 0 交易的期权在年度、月度、周度、日度、秒度的不同层次、不同深度数据，可以满足数据“大”的标准。而“人工智能”从 1956 年 Dartmouth 学会上提出至今已经满了一个 60 年，其研究、开发用于模拟、延伸和扩展人的智能的理论、方法、技术及应用系统的技术科学，具体研究包括机器人、语言识别、图像识别、自然语言处理和专家系统等，其中的核心是机器学习。机器学习设计和分析这些

*　本文系国家自然科学基金重点项目“中国金融体系的演化规律和变革管理”(项目号: 71733004) 的阶段性成果。

让计算机可以自动“学习”的算法，正是期权策略中，对波动率预测可以使用的新方法。

因此，利用交易数据和算法人工智能，将机器学习技术应用于期权金融市场，提高期权风险管理水平和投资决策效率，是本文尝试的一个方向。本文主要目的是探索机器学习在期权波动预测中的应用，主要创新是提出波动率预测“SKRG 递进集成”法，较高质量预测了隐含波动率，该机器学习有利于提高波动率预测的精度。具体而言，SKRG 递进集成法，是基于中国波指预测的特征(Feature)，分别运用随机森林、GBM 及 K 临近等算法，搭建了层层递进的 48 个指标，并在逻辑上做集成处理，得到最优化成果。

文献综述

对收益波动率的建模和预测是金融市场研究的一个重要议题。主流的方法是通过历史数据即时间序列模型。

Engle 等较早提出 ARCH 类模型，之后学者提出 GARCH 等一系列修正模型。黄海南等(2007)运用 GARCH 模型对上证指数收益率进行估计及样本外预测，然后以已实现波动率作为波动率预测的评价标准，通过 M－Z 回归和损失函数来评价 GARCH 类模型的波动率预测表现。结果表明，无论是样本内还是样本外，GARCH 类模型都能够较好地预测上证指数的收益波动率。其中，偏斜 t－分布假设下的 GJR(1,1)模型的预测能力最强。赵华等(2011)分别基于误差项服从正态分布、t 分布、广义误差分布的 GARCH 族模型和 MRS－GARCH 模型对中国股市波动的结构变化特征进行实证研究。结果表明，中国股市存在显著的高、低波动状态，MRS－GARCH 模型预测效果总体上优于 GARCH 族模型。李汉东等(2003)讨论了在金融时间序列中广泛应用的两类波动性模型，即自回归条件异方差(ARCH)模型和随机波动(SV)模型的关系问题，认为一个离散的 EGARCH(1,1)模型在弱 GARCH 过程的条件下与一个离散的 SV 模型是一一对应的。在此基础上进一步讨论了 EGARCH(1,1)模型和 SV 模型的单位根问题，结果表明：两类模型的单位根存在对应的关系，即二者的持续性能够通过随机微分方程的形式来传递。但 GARCH 模型的缺点在于，无法考虑期权波动率二阶的复杂性和非线性特征。

部分学者利用贝叶斯原理对随机波动率模型进行研究。Jacquieret al.(2002)利用股票的收益率和换手率的日数据和周数据，通过抽样实验来比较贝叶斯估计法、矩量法和拟极大似然法。实验结果表明：在参数估计，贝叶斯估计法要优于另外两种方法。蒋祥林等(2005)基于贝叶斯原理对随机波动性模型进行研究，并将随机波动率模型应用于股市风险价值的估计与预测。针对中国股市数据进行的实证结果表明：与 GARCH 模型相比，随机波动率模型能更好地描述股票市场回报的异方差和波动率的序列相关性，基于随机波动率的 VaR 较 GARCH 模型的 VaR 具有更高的精度。类似地，罗嘉雯等(2017)通过构建包含时变系数和动态方差的贝叶斯 HAR 潜在因子模型，对我国金融期货的高频已实现波动率进行预测。结果表明，时变贝叶斯潜在因子模型在所有参与比较的预测模型当中具有最优的短期、中期和长期预测效果。同时，在股指期货和国债期货的预测模型中加入投机活动变量可以获得更好的预测效果。但贝叶斯估计法难以处理期权的不同执行价、不同到期日、不同执行权的欧式或美式等多维度特征，常常依赖于单因素的分布条件。

陈蓉等(2010)利用香港恒生指数期权的数据，对隐含波动率曲面动态过程进行建模和估计，建立了一个五因子随机隐含波动率模型。在模型的估计方法上，首次引入了基于小样本面板数据的扩展的卡尔曼滤波法。结果显示，在香港市场上，扩展的卡尔曼滤波法比传统的两步法可以得到更好的估计结果，五因子随机隐含波动率模型能很好地刻画恒指期权隐含波动率曲面的变动规律，效果明显优于静态隐含波动率模型。但中国市场的期权交易尚不充分活跃的情形下，部分非主力合约的波动率曲面的建立容易失真。

除了传统的波动率预测模型之外，部分学者不断提出新的预测模型。魏宇等(2015)在已有的多分形波动率(multifractal volatility)测度方法的基础上提出新的波动率测度方法及模型。基于上证综指的 5 min 高频数据，发现不论是短记忆模型还是长记忆模型，多分形波动率模型的预测精度明显优于 GARCH 族模型，且长记忆模型的预测能力要好于短记忆模型。郑振龙等(2017)

根据新的隐含波动率半参数模型，利用 MATLAB 编程，选择香港小型恒生指数期权 2013 年 1 月到 2015 年 3 月的日交易数据，分别实现了滚动加权平均法与 BP 神经网络法对参数的周期性时间序列进行外推预测，发现 BP 神经网络法明显优于滚动加权平均法。这些尝试是机器学习在期权波动率预测的尝试，尽管主要局限于上证股票指数或香港期权市场。

近年来机器学习在金融市场预测中得到越来越多的应用。Rose(2013) 将机器学习用于流行病学研究，结果发现超级学习者在预测死亡率方面比单一算法具有优势。李光明(2013) 基于粗糙集的神经网络模型，针对国有企业目前的经营绩效进行分类，实验结果显示约简后的国有资产指标集可以很好地反映国有企业的财务风险情况。彭岩等(2017) 讨论了基于案例的推理(CBR，Case based Reasoning) 、支持向量机(SVM，SupportVector Machine) 以及人工神经网络(ANN，Artificial Neural Network) 等机器学习方法在风险预测中的作用。曹正凤(2014) 通过比较分析价值策略和成长策略，提出以价值成长投资策略(GARP) 理念为基础的选股模型指标体系，通过样本数据发现，使用随机森林算法可以更好地完成股票分类，实现更好收益。辛治运和顾明(2008) 基于最小二乘支持向量机的对复杂金融时间序列进行预测，吴微等(2001) 运用 BP 神经网络预测股票市场涨跌，张炜等(2015) 基于自适应遗传算法对股票未来走势进行预测，苏治等(2013) 通过核主成分遗传算法对 SVR 选股模型进行改进，王梦雪(2016) 利用拍拍贷平台的借贷数据，通过各种机器学习的算法选择风控模型的因子，并对约定的违约进行预测，得到比较满意的结果。整体上看，这些研究标的物多为股票或借贷，在国内的金融期权上尚属于空白。

通过上述文献可以看出，尽管机器学习正越来越多地用于金融预测与风险管理，但用于期权风险预警、预测波动率的文献还较少。同时，如何在期权隐含波动率预测上建立一个机器学习应用模型，这一空白需要填补。因此，本文运用机器算法机制，综合随机森林、GBM 及 K 临近等算法，提出“SKRG 递进集成”法模型，用于期权风险预警，并通过实盘数据进行了有效检验。

基于机器学习算法的期权波动率预测

(一) 机器学习在期权波动率预测上的评价标准

能否高质量地评价机器学习方法对波动的预测，需要建立科学的评价指标。根据机器学习的实际应用情况，机器学习一般分为三类: 监督学习(Supervised Learning，SL) ，非监督学习(Unsupervised learning，UL) ，和强化学习(Reinforcement Learning，RL) 。本文应用监督学习可判别预测的效果。监督学习是在给定训练样本，该样本既有数据，又有数据对应结果，利用该样本进行训练得到模型，然后利用该模型，将所有的输入映射为相应的输出，之后对输出进行简单的判断，从而达到分类或回归的过程。因而监督学习是原始数据中既有特征值，也有标签值的机器学习。

因此，本文机器学习的主要评价指标包括四个方面，如下图 1 所示: (1) 准确率(Accuracy) ，指对于给定的测试数据集，分类器正确分类的样本数和总样本数之比; (2) 精确率(Precision) ，每次预测成功的概率; (3) 召回率(Recall) ，反映的是能够识别风险的概率; (4) F1 - Score，指精确率和召回率的调和均值。

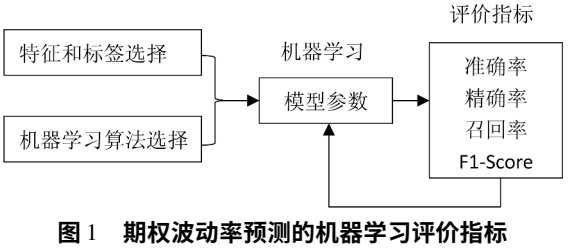

图 1　期权波动率预测的机器学习评价指标

由图 1 可知，本文在机器学习模型效果上，注重四个指标，分别是预测的准确率、精确率、召回率和二者的调和均值。通过四方面的对比，寻找较优的预测模型。

(二) 期权波动率预测特征(Feature) 与标签(Label) 选择

在卖出类期权类策略中，期权的价值表示为:

$$\frac{1}{2}(\sigma^2 - \tilde{\sigma}^2)\int_{t_0}^{T} e^{-r(t-t_0)} S^2 \Gamma^i dt.$$

由于 Vega 为负，如果隐含波动率大幅上涨，势必带来较大的投资损失。因此，我们把波动率变化幅度予以分类，根据 Scott Mixon(2007) 的分

类法，本文把波动幅度在2%以内定义为安全类，把超过2%定义为风险类。

对于隐含波动率的标的选择，本文选择中国波指，000188. SH，其特点是构造较公允、波动价格的跟踪误差较小、能够较好反映期权的隐含波动状况，反映市场情绪。

对于训练和测试的时间段的选择中，依据交易量较大的2015年2月9日至2017年10月18日，共655个交易日。

在隐含波动率的因子选择，由于隐含波动率的上涨下跌与标的资产实际的波动状况以及市场的情绪有关，考虑到数据的可得性，本文选取实际波动状况、历史波动率、与波动状况相关的技术指标、波动率预测以及期权市场数据五大类数据，共48个相关指标作为隐含波动率的影响因素。这些因子基本覆盖了期权理论因素点或各大历史文献研究的主要指标，具体如下表1所示。

表1　期权隐含波动率的影响因子选择

实际波动状况	涨跌幅，振幅，成交量，前一日路径长度，5日路径长度EMA值，10日路径长度EMA值，30日路径长度EMA值。前一日极差，10日极差EMA值，30日极差EMA值。
历史波动率类	10日收盘价波动率，30日收盘价波动率，60日收盘价波动率，10日Parkinson波动率，30日Parkinson波动率，60日Parkinson波动率，10日比值，30日比值，60日比值。
与波动相关技术指标	MTM，RSI，VSTD，VOSC，WVAD，SI，SOBV，VR，3日ATR，7日ATR，14日ATR。
波动率预测	GRACH预测：10日波动率，30日波动率，60日波动率。 CARR预测：10日波动率，30日波动率，60日波动率。
期权市场数据	P/C成交量，P/C持仓量，P/C5日成交量均值，P/5日C持仓量，前一日隐含波动率。

由表1可知，期权隐含波动率的影响因子中，包括实际波动状况，其可以细分为涨跌幅、成交量、振幅等指标，也包括历史波动率指标，不同日期的收盘价波动率或Parkinson指标，以及各类call与put的比值等。

（三）期权波动率机器学习算法模型构建

基于前述算法，本文开始通过数据对模型进行训练，优化模型参数。在训练的过程中，依据较高的“精准率”，提升“召回率”逐步优化模型。机器学习的算法中，考虑到因子数据量大、维度较高，选择先用降维映射的算法，因此首先选择SVM算法。同时，SVM可以克服因变量数据较小的不足。

1. SVM算法降维分类

SVM即支持向量机，这是一种监督学习方法，主要用于分析数据、识别模式，对数据的分类分析和回归分析①。由于支持向量机可以将分类问题转化为一个不等式约束下的二次规划问题，并用核函数代替向高维空间的非线性映射，因而较好地解决了高维数问题，成为现阶段统计理论发展最快的研究方向之一。鉴于我们的数据样本数量只有655份，属于小样本数据集，而SVM在小样本数据上有较为优秀的表现，因此先使用支持向量机对风险预警问题进行处理。

由于我们的数据维度较高，因此需要用RBF核函数将样本映射到高维空间，在参数的训练过程中我们主要训练两个参数，一个是gamma，是RBF函数自带的一个参数。gamma越大，支持向量越少，gamma值越小，支持向量越多。我们调整gamma的值在0.01至1.5的范围内，其精确率、召回率以及F1值有如下变化（图2）。

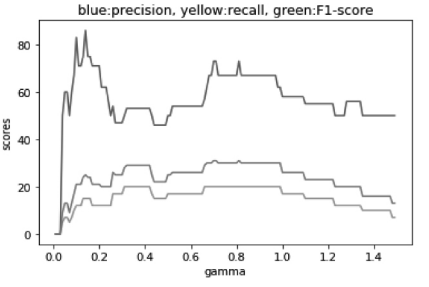

图2　gamma值变动时指标曲线图

我们可以看到在gamma在0.8左右有着较好的性能，且鲁棒性较好。另一个是惩罚系数C，即对误差的宽容度。C越高，说明越不能容忍出现误差，容易过拟合。C越小，容易欠拟合。C过大或过小，泛化能力变差。我们调整惩罚系数C的值在1至5的范围内，其精确率、召回率以及F1值变化如图3。

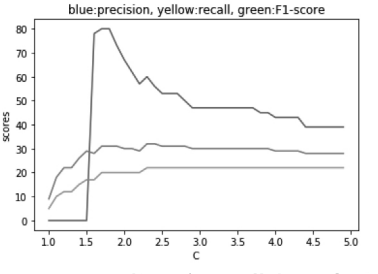

图3　惩罚系数C变动时指标曲线图

当惩罚系数C小于1.5时预测的精准度是很

低的，在 1.5 到 2 之间有一个较高值，之后逐渐衰减，综合考虑我们选择 C 的值为 1.8。通过调参后，支持向量机在测试集上的表现如下(图 4) 。

accuracy 0.822335025381				
	precision	recall	f1-score	support
Benign	0.82	0.99	0.90	156
Malignant	0.80	0.20	0.31	41
avg / total	0.82	0.82	0.78	197

图 4　SVM 在测试集上的表现结果

由图 4 可见，SVM 具有较好效果，精准率可以达到 0.8，召回率也在 0.8 左右。但在实际交易中，考虑到我们更关心波动率较大的突变，而不是每次均等变化，前文中的“风险类”样本，是我们更关注的对象。因此我们用 KNN 进行优化。

2. KNN 优化样本的不平衡

由于我们的数据存在样本不平衡的现象，“风险类”的样本明显少于“安全类”。为有效解决样本不平衡的问题，我们将训练 KNN 模型来对问题进行处理。经过数据处理后我们开始对模型进行参数调节，由于 KNN 算法是一种被动的算法，没有一个训练的过程，因此我们在训练集内部做十折交叉验证来选取一个合适的 k 值以及加权方式。其精准率的展示如下图 5、图 6。

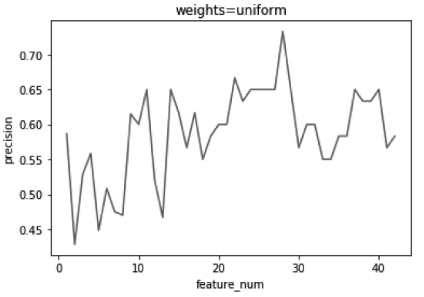

图 5　等权重时下精准率与 k 值的关系图

图 5 表示当取各个数据点权重相等时，k 的不同取值对精准率的影响，图 6 表示当给各个数据点按距离分之加权时，k 的不同取值对精准率的影响。通过两幅图的对比我们可以发现，对各个数据点赋予相等权重的效果明显要更好一些。同时发现当 k 值在 20 到 30 之间有着较好的效果。通过调参后，KNN 算法在测试集上的表现如下(图 7) 。通过图 7 可以看出，KNN 算法在精准率上的表现和随机森林相同，但是在召回率上要更好一些。

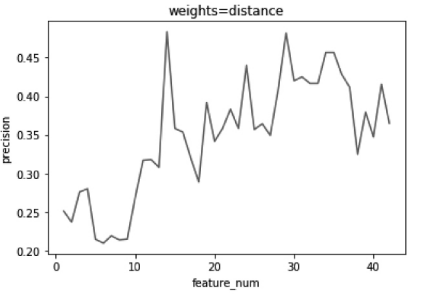

图 6　加权后精准率与 k 值的关系图

accracy: 0.822335025381				
	precision	recall	f1-score	support
Benign	0.83	0.98	0.90	156
Malignant	0.75	0.22	0.34	41
avg / total	0.81	0.82	0.78	197

图 7　KNN 在测试集上的表现结果

由图 7 可以看出，KNN 算法在精准率上的表现和 SVM 相近，但在召回率上更好一些。

3. 在 SVM 和 KNN 上用 RF 优化特征值权重

无论是支持向量机还是 KNN 算法都是同时对多组数据进行分析处理，虽然我们提前会对特征做一些筛选工作，排除一些相关性较差的特征，但在留下的特征当中仍是赋予了相同的权重，而实际上每个特征对隐含波动率的影响不会是完全相同的。而树模型是每次只对单个特征进行处理，每次都会选择信息增益最大的特征作为判断模块建立子结点，当节点内的样本全部归为一类或是到达我们规定的深度便会停止继续划分，这样可以使得我们根据特征的重要程度依次对特征进行处理。基于这个特点我们进一步使用随机森林对问题进行处理。

最大特征数(Max_Features) 是指随机森林允许单个决策树使用特征的最大数量。增加最大特征数一般能提高模型的性能，因为在每个节点上，我们有更多的选择可以考虑。然而这未必完全是对的，因为它降低了单个树的多样性，而这正是随机森林独特的优点。但是可以肯定的是，通过增加最大特征数会降低算法的速度。因此需要适当的平衡和选择最佳最大特征数。为此我们调节最大特征数的取值 0 到 40，其精确率、召回率以及 F1 值有如下变化(图 8、图 9) 。

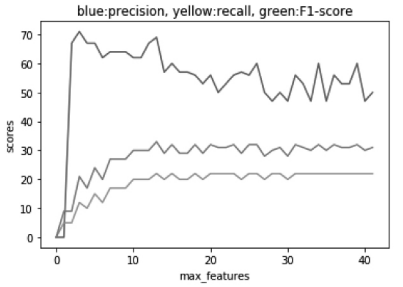

图 8　最大特征数与评价指标关系图

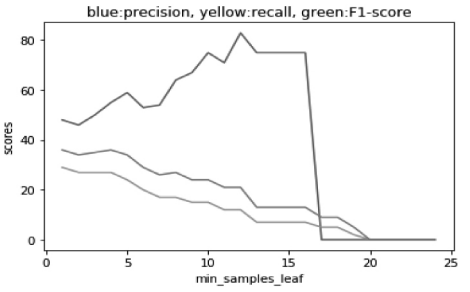

图 9　最小叶子样本数与评价指标关系图

从图 8 可以看到，在最大特征数非常小的时候，模型基本没有什么预测能力，三个值都非常的低，最大特征数取 3 到 10 的区间范围时，精确率较高，召回率及 F1 值较低且有逐渐上升的趋势，当最大特征数大于 10 之后，精确率有稍微下降的趋势，且召回率和 F1 值逐渐趋于稳定。综合考虑，我们取最大特征数的值为 10。

最小叶子样本数(min_sample_leaf) 控制着树枝在分叉时的最小样本数，当前节点样本数小于这个值的时候，当前节点停止构建，作为决策树的叶子节点。这个值决定着决策树的深度，一般而言取值越小性能会越好，但如果叶子太小会使模型更容易捕捉训练数据中的噪声，使得决策树较为容易过拟合。我们调节最大特征数的取值 0 到 40，其精确率，召回率以及 F1 值变化如图 9。

我们看到当取值越小时，召回率越高，取值越大，召回率越低，主要原因是我们的数据有一定的偏态，归为“安全类”的数据大约占到了 77%，树模型的深度越低，越容易被归为“安全类”，当取值为 10 到 15 时，精准率有一个较高的取值。综合考虑，我们取最小叶子样本数的值为 11。通过参数调节后随机森林模型的性能如下(见图 10) :

```
accuracy: 0.812182741117
             precision    recall  f1-score   support

     Benign       0.81      0.99      0.89       156
  Malignant       0.75      0.15      0.24        41

avg / total       0.80      0.81      0.76       197
```

图 10　随机森林在测试集上的表现结果

4. 考虑样本不平衡和权重差以后的 GBDT 梯度提升

在测试上述集中共有 41 个风险类，随机森林模型可以识别出其中的 15%，其预测的精准率达到 75%。但是召回率要略低于 KNN 算法。在随机森林中使用的是 Bagging 的方法，每轮抽取的训练集的选择是随机的，各轮训练集之间相互独立，各个预测函数没有权重。相比于 bagging，在集成树模型中还有一种 boosting 方法，在开始时会给每个样本相等的权重，然后用该算法对训练集训练 n 轮，每轮训练后，会对训练错的样本加大权重，也就是让学习算法在后续的学习中集中对比较难的训练例进行学习，从而得到一个预测函数序列，其中预测函数也有一定的权重，预测效果好的预测函数权重较大，反之较小。Bagging 采用均匀取样，而 boosting 根据错误率来取样，因此 boosting 的分类精度要优于 bagging，梯度提升决策树是一种使用 boosting 的方法，在这一部分我们将使用梯度决策树算法来对问题进行处理。

与随机森林类似，梯度提升决策树也是以决策树作为基础分类器的一种集成模型，因此它也存在决策树中的一些参数，例如最小叶子样本数、最大深度等，但它同时也包含了调节模型中 boosting 操作的参数以及调节模型总体各项运作的参数。下面通过实证分析考察子样本数、学习率、最大特征数以及最小叶子样本数对模型性能的影响，并确定最佳模型参数。

实际中，子样本数是指每棵决策树中所包含的全体样本的数量，一般这个值选取的越大，会使得单棵树中获取的信息量也越大，性能也越高，但同时也会造成树与树之间差异性的减小，容易造成过拟合。图 11 反映了当子样本数变化时各指标的状况，从图中我们可以看到当子样本数取 30% 到 50% 时，精准率与召回率都有着较好的表现。

设定了初始的权重值之后，每一次树分类都会更新这个值，而 learning rate 控制着每次更新的

幅度。一般来说这个值不应该设得比较大，因为较小的 learning rate 使得模型对不同的树更加稳健，能更好地综合它们的结果。当然我们也需要考虑到运算效率，学习率设置得越小，运算量越大，在可接受的运算量范围内，我们可以尽量地设置较小的学习率。图 12 反映了学习率变化时各指标的状况，从图中我们可以看到较小的学习率确实有助于提高精准率。

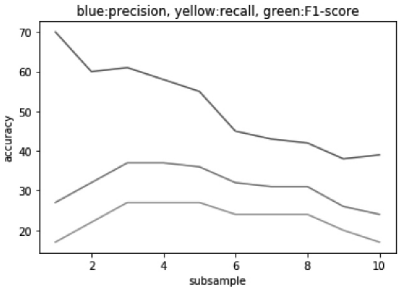

图 11　子样本数与评价指标关系图

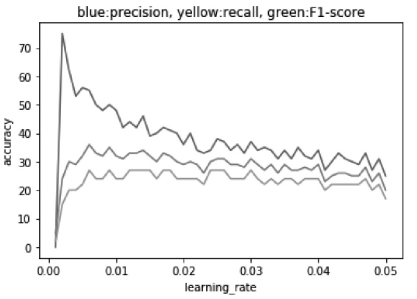

图 12　学习率与评价指标关系图

与随机森林模型相同，我们同样对最小叶子样本数以及最大特征数进行参数调整，各指标表现如图 13、图 14。图 13 表现的是不同最小叶子节点对指标的影响，可以看到在取值为 20 左右的时候，精准率有着将近 80% 的优异表现，同时召回率也不是特别的低，图 14 展现的是不同的最大特征值对指标的影响，可以看到在取值为 10 到 20 之间时，精准率有着较为优异的表现。

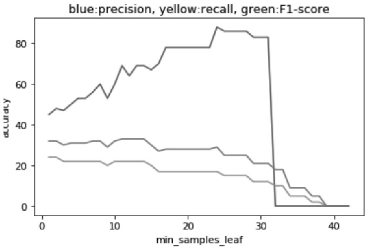

图 13　最大特征数与评价指标关系图

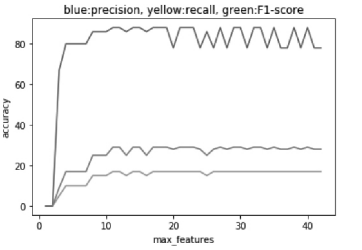

图 14　最小叶子样本数与评价指标关系图

通过参数调节后梯度提升决策树模型的性能如下：

accurary: 0.832487309645

	precision	recall	f1-score	support
Benign	0.83	0.99	0.90	156
Malignant	0.90	0.22	0.35	41
avg / total	0.84	0.83	0.79	197

图 15　GBDT 在测试集上的表现结果

在测试集中共有 41 个风险类，梯度提升决策树模型可以识别出其中的 22%，其预测的精准率达到 90%，整体表现继续得到优化和提升。

5. 整体算法递进集成

上述 SVM、KNN、RF、GBDT 在期权波动率上的四步预测，我们简称为 SKRG 算法纵向集成。从逐步算法结果来看，集成效果较好。除了这种纵向层层递进式算法调仓，我们尝试把四个预测模型的预测结果取“或”，也就是说只要有一个模型发出预警信号时，我们即认为第二天隐波会上涨 2%，在测试集上的横向集成表现如下(图 16) 。

	precision	recall	f1-score	support
Benign	0.84	0.98	0.91	156
Malignant	0.80	0.29	0.43	41
avg / total	0.83	0.84	0.81	197

图 16　四个模型集成后在测试集上的表现结果

整体来看，相比于单个模型的表现，横向集成后提高了召回率，我们可以预测出将近三分之一的风险，不过同样也把预测的准确度降到了 80%。跟单个模型比起来只是会好于随机森林，跟其他三个模型相比效果都要差一些。单从预测效果上来说，随机森林方法表现最差，由于其把集成模型的精准率拉低，我们排除掉它，只利用其他三个模型在测试集上进行预测，表现如下：

	precision	recall	f1-score	support
Benign	0.84	0.99	0.91	156
Malignant	0.86	0.29	0.44	41
avg / total	0.84	0.84	0.81	197

图 17　SVM、KNN 与 GBDT 三个模型集成后在测试集上的表现结果

可以看到，在召回率没有下降的情况下，精准率得到了提升，说明随机森林可以预测出来的风险都被其他三个模型覆盖掉，因此我们在集合模型中只选择支持向量机、KNN 和梯度提升决策树三个模型作为基础模型。同样我们把集成模型来预测样本外的数据，我们取 2017 年 10 月 19 日至 2018 年 2 月 6 日的数据进行预测，其表现如下（图 18）：

	precision	recall	f1-score	support
Benign	0.63	1.00	0.77	46
Malignant	1.00	0.18	0.31	33
avg / total	0.78	0.66	0.58	79

图 18　SVM、KNN 与 GBDT 三个模型集成后在样本外数据集上的表现结果

我们可以看到在这段时间里，模型的精准率达到了 100%，也就是说在这段时间里每当模型发出风险预警时，都没有发生误报的状况，相比而言召回率为 18%，也就是说在发生风险的 33 天里，我们总共预测出了 6 次。相比于在测试集中的表现，在样本外有着更高的精准率以及较低的召回率。

SKRG 递进集成算法与传统预测方法的比较

总体而言，期权波动率预测的机器学习算法中，由于期权波动率的因子数据量较大，维度较高，选择先用降维映射的 SVM 算法，但 SVM 不会考虑“风险”样本的特殊性，因此增加 KNN 的优化。又由于 SVM 和 KNN 都隐含样本权重相等，需要调整考虑特征值情况，因此引入 RF，并精细化地提升梯度引入 GBDT 和纵向、横向集成，这一过程我们称为 SKRG 递进集成期权隐波机器学习算法。

在波动率预测的问题上，较为流行的方法是利用 GARCH 模型来进行预测，GARCH 模型是由 Bollerslev 在 1986 年提出的，他在原自回归条件异方差模型进行改进，提化了该模型，该模型在一定程度上解决了待估参数不断增加从而增大求解难度，以及导致解释变量容易引发多重共线性问题。运用 GARCH(1,1) 来对隐含波动率进行预测，在 2015 年 2 月 9 日至 2017 年 10 月 18 日的 655 个交易日里，其表现如下：

	precision	recall	f1-score	support
Benign	0.77	0.92	0.84	506
Malignant	0.22	0.08	0.12	149
avg / total	0.65	0.73	0.67	655

图 19　GARCH 模型在样本集与数据集上的表现结果

从图 19 可以看到其精准率只有 22%，召回率只有 8%，都远远低于我们利用机器学习的预测能力。原因在于：GARCH 模型仅仅利用到了过去 n 个交易日的收益率、方差以及长期均方差这几项历史数据，而隐含波动率作为衡量期权价格的指标，反映了投资者对市场情绪的预期，绝不仅仅是这两三个因子可以刻画出来的。机器学习模型可以同时处理几十个维度的数据，更为全面的多角度的对隐波的涨跌去进行思考判断，同时利用了多个模型的差异性，相当于让多个专家来共同进行抉择判断，相对而言会有更强的预测能力。

结　论

基于期权波动率传统预测方法的不足，我们将机器学习算法引入到预测模型中。考虑期权隐波预测的高维度数据难度与特征值情况，依次引入过 SVM 机器学习、KNN 样本不平衡机器学习、RF 划分、GBDT 优化、算法递进集成完成机器学习建模过程。结果显示，SKRG 的预测效果好于传统的 GARCH 模型。SKRG 丰富了期权随机波动率预测领域的相关文献，为期权风险预警提供了新的方法。

①原始的支持向量机算法由 Vladimir Vapnik 发明，而当前的标准化由 Corinna Cortes 和 Vladimir Vapnik 提出。

参考文献

1. 张炜、范年柏、汪文佳《基于自适应遗传算法的股票预测模型研究》，《计算机工程与应用》2015 年第 4 期。
2. 辛治运、顾明《基于最小二乘支持向量机的复杂金融时间序列预测》，《清华大学学报》(自然科学版) 2008 年第 7 期。
3. 苏治、傅晓媛《核主成分遗传算法与 SVR 选股模型改进》，《统计研究》2013 年第 5 期。

4. 彭丽芳、孟志青、姜华、田密《基于时间序列的支持向量机在股票预测中的应用》,《计算技术与自动化》2006 年第 3 期。
5. 郑振龙、黄薏舟《波动率预测: GARCH 模型与隐含波动率》,《数量经济技术经济研究》2010 年第 1 期。
6. 屈满学、王鹏飞《我国波动率指数预测能力研究——基于隐含波动率的信息比较》,《经济问题》2017 年第 1 期。
7. 魏宇、马锋、黄登仕《多分形波动率预测模型及其 MCS 检验》,《管理科学学报》2015 年第 8 期。
8. 罗嘉雯、陈浪南《基于贝叶斯因子模型金融高频波动率预测研究》,《管理科学学报》2017 年第 8 期。
9. 赵华、蔡建文《基于 MRS - GARCH 模型的中国股市波动率估计与预测》,《数理统计与管理》2011 年第 5 期。
10. 蒋祥林、王春峰《基于贝叶斯原理的随机波动率模型分析及其应用》,《系统工程》2005 年第 10 期。
11. 马俊美、杨宇婷、顾桂定等《随机波动率模型下基于精确模拟算法的期权计算理论》,《同济大学学报》(自然科学版) 2017 年第 10 期。
12. 陈蓉、吕恺《隐含波动率曲面: 建模与实证》,《金融研究》2010 年第 8 期。
13. 黄海南、钟伟《GARCH 类模型波动率预测评价》,《中国管理科学》2007 年第 6 期。
14. 张彩玉、屠巧平《期权风险管理探讨》,《商丘师范学院学报》2006 年第 4 期。
15. 李时运《新形势下期货期权风险管理研究》,《经贸实践》2015 年第 8 期。
16. 何问陶、徐华云《期权定价理论在风险管理中的应用》,《商业研究》2004 年第 13 期。
17. 陈荣达《基于 Delta - Gamma - Theta 模型的外汇期权风险度量》,《系统工程理论与实践》2005 年第 7 期。
18. 刘国新、梁靖廷、潘祥杰《基于期权理论的风险投资决策分析》,《价值工程》2002 年第 6 期。
19. 李书言《上证 50ETF 期权风险管理研究》,东北师范大学,2016 年。
20. 赵建、薛奕达《基于波动率指数的期权对冲策略研究》,《河北工业科技》2009 年第 6 期。
21. 潘碧云《期权风险的对冲策略分析》,《科技信息》2013 年第 3 期。
22. 李倩、张潇尹《ETF 期权的交易风险及对策研究——以上证 50ETF 期权为例》,《沈阳工业大学学报》(社会科学版) 2016 年第 3 期。
23. 李光明《基于机器学习的国有资产监管系统风险预警模型的研究》,重庆理工大学,2013 年。
24. Black F, Scholes M. The Pricing of Options and Corporate Liabilities, *Journal of Political Economy*, 1973, 81(3): 637 - 654.
25. Mixon Scott, Option markets and implied volatility: Past versus present, *Journal of Financial Economics*, 2009, 94(2): 171 - 191.
26. Jacquier Eric, Nicholas G Polson & Peter E Rossi. Bayesian Analysis of Stochastic Volatility Models, *Journal of Business & Economic Statistics*, 2002,20(1): 69 - 87.
27. Sílvia Gonalves and Massimo Guidolin. Predictable Dynamics in the S&P 500 Index Options Implied Volatility Surface, *The Journal of Business*, 2006, 79(3): 1591 - 1635.
28. Rose Sherri, Mortality Risk Score Prediction in an Elderly Population Using Machine Learning, *American Journal of Epidemiology*, 2013,177(5): 443 - 452
29. Tortelli R, Ruggieri M, Cortese R, et al. Option Pricing: A Simplified Approach., *Journal of Financial Economics*, 1979, 7(3): 229 - 263.
30. Frank J. Fabozzi, TommasoPaletta, Silvia Stanescu, RaduTunaru. An improved method for pricing and hedging long dated American options European Journal of Operational Research, 2016,254(2): 656 - 666.
31. Jianfeng Hu, Does option trading convey stock price information? *Journal of Financial Economics*, 2014, 111(3): 625 - 645.
32. Zhang J, Ida M Friberg, Ann Kift - Morgan. Machine - learning algorithms define pathogen - specific local immune fingerprints in peritoneal dialysis patients with bacterial infections:, *Kidney International*, 2017, 92(1): 179 - 191.
33. Aziz O, Musngi M, Park E J. A comparison of accuracy of fall detection algorithms (threshold - based vs. machine learning) using waist - mounted tri - axial accelerometer signals from a comprehensive set of falls and non - fall trials, *Medical & Biological Engineering & Computing*, 2017, 55(1): 45 - 55.
34. Zhu D M, Lu J, Ching W K. Option Pricing Under a Stochastic Interest Rate and Volatility Model with Hidden Markovian Regime - Switching, *Computational Economics*, 2017(11): 1 - 32.
35. Imai J. Dimension Reduction for Pricing Options Under Multidimensional LévyProcesses, *Asia - Pacific Financial Markets*, 2014, 22(1): 1 - 26.
36. Wang C W, Yang S S, Huang J W. Analytic option pricing and risk measures under a regime - switching generalized hyperbolic model with an application to equity - linked insurance, *Quantitative Finance*, 2017, 17(22): 1 - 15.

作者简介: 马天平,对外经济贸易大学博士后; 吴卫星,对外经济贸易大学金融系教授。北京,100032

〔责任编辑: 凌　羽〕