【广发固收刘郁团队】如何利用机器学习方法构建转债择券模型——转债入门手册之六

原创 刘郁 田乐蒙 郁言债市 *4天前*

来自专辑

【广发固收】转债系列入门手册

郁言债市微信小程序

区域经济、城投、地产数据一网打尽

欢迎关注使用！

摘 要

我们在此前的报告《转债定价方法进化史——转债入门手册之四》中曾提到，由于可转债是一种构成极为复杂的衍生品，基于B-S期权定价思想的主流资产定价方法，在国内市场期权定价问题中的适应性存在较大问题；另一方面，各条款的路径依赖特征和相互关联也为转债整体定价增添了难度。

对于可转债这样一个复杂的产品，机器学习方法的"黑箱"特性反而成为了能有效避免方法陷入严苛假定约束的新路径。机器学习方法的思想是不断利用经验数据对数理模型的判断效力进行训练和优化，通过学习算法（learning algorithm）使其"积累经验"，最后使得模型可以尽可能准确地模仿人脑，完成基于历史数据的经验判断。

在本篇报告中，我们选取了三种常见的学习方法来演示构建机器学习择券体系的思路和流程，分别是**随机森林（Random Forest）、XGBoosting**和**支持向量机。**

随机森林方法是一套在决策树方法基础上，基于集成学习思想发展出的方法体系。随机森林在自助采样原理的基础上，对集成学习中的Bagging方法进行了拓展，生成多个相对判断效力较弱的决策树，形成一个更为强效的"森林"，从而对研究目标问题进行更为准确的判别。

Boosting方法与上一节中介绍的随机森林方法一样，同样属于常见的集成学习方法，即核心是将多个弱学习器进行联合后，将其增强为一个更强的学习器，原理较随机森林通常更为简单。本篇报告中实际选用的XGBoosting方法是一种基于GBDT的改进方法。

支持向量机的思想就是在各类样本空间中，利用各种手段寻找能有效将其划分的超平面，在得到划分超平面之后，对于新的样本，我们便可以通过定位其与划分超平面的相对位置，来对其所属的类别进行判断。

为探究机器学习方法在转债择券研究中的实际效果，我们运用上述三种方法分别建立了月度换仓的等权择券模型，并利用国内转债市场的历史数据对三种方法的择券效果进行了初步测试。

从测试的结果来看，三种方法均在样本测试区间取得了超过中证转债指数的累计收益。其中，基于机器学习技术构建的择券模型都表现出了超过指数的累计收益，其中随机森林方法在稳定性和运算效率方面占据显著优势，XGBoosting在累计收益水平和判断准确率方面占优，而SVM模型的表现则略逊于随机森林和XGBoosting。

核心假设风险。转债交易和发行相关规则出现重大调整

1

为何要在转债研究中尝试引入机器学习方法?

MLF超量续作，背后是超储率降至2017年水平MLF超量续作，背后是超储率降至2017年水平MLF超量续作，背后是超储率降至2017年水平

关于什么是机器学习方法，我们可以举一个简单的例子：当我们初次接触到某一只陌生的转债新券时，在其发行文件中，我们注意到这只新券的正股具有极高的关注度，并且发行规模大，主体信用评级较高，此时我们会下意识地感觉到，这只新券的上市价格大概率不会便宜。例如，在此前国内火锅料龙头安井食品的第二期转债——安20转债临近上市前，恐怕很少有经验丰富的投研人员会认为其上市时价格会处在较低区间。**由此，即便在新券尚未上市时，我们仍可以在脑中对其上市价格形成一个下意识的判断。这是因为，在我们长期以来跟踪新券发行上市的过程中，已经对"不同特点新券的上市价格高低"这个问题积累了大量经验，并在历史数据的基础上训练出了一套大致预判转债上市价格的经验法则，所以当我们在看到一些代表性的信号时（例如前文中提到的正股关注度、规模、等级），便可以提前对其可能的上市价格进行预判。**

**
**

在上述的例子中，投研人员的经验在预判过程中起到了决定性的作用，然而由于人脑的"硬件条件"限制，即便是最为经验老到的研究者，在上述的判断过程中也很可能由于主观认知偏差、记忆力限制等方面的掣肘，做出与历史规律并不相符的判断，无法完全发挥出历史数据中蕴含的信息。然而，如果我们能构建一套行之有效的量化方法，借助计算机强大的运算能力和高度的纪律性、客观性，替代人脑完成这一过程，便可以在最大限度上发挥历史数据的威力，并或许能得到比经验丰富的投研人员更为有效的预判结果。而机器学习方法，正是这样一套通过数量化手段构建的方法体系，其思想是不断利用经验数据对数理模型的判断效力进行训练和优化，通过学习算法（learning algorithm）使其"积累经验"，最后使得模型可以尽可能准确地模仿人脑完成上述判断过程的方法体系。[1]

在前文的例子中，如果用机器学习的语言来表达，我们在上述例子中提到的新券信息——正股关注度、规模、等级等通常被称为**属性（Attribute）**；而这些信息上的取值——正股关注度火热[2]、募集规模50亿元、主体信用评级AA+等，则通常被称作**属性值（Attribute Value）**。在前文的例子中，我们观测到的属性信息为：（正股关注度=火热；规模=50亿元；主体信用评级=AA+），这样的一组信息便构成了一个**样本（Sample）**，而如果把自己长期以来参与新券价格分析的经验进行汇总，便可以得到大量的经验样本，而这样的一组样本汇总被称作**数据集（Data Set）**。最后，为了使得我们的方法可以用于判别，我们还需要在数据集中加入其重要的结论信息，就是每一只新券的实际上市价格——到底是处在高、中、还是低区间（例如将100元以下视为低、100-115元为中等、115元以上为高）。如实际上市价格这样的结果数据通常被称作**标记（label）**，加入了标记的样本被称作**样例（example）**。在这样的一个例子当中，我们便可以利用具有上述要素的数据集来对我们的梳理方法进行训练，使其积累经验，完成学习任务。

表1：大量新券属性和上市价格信息可以形成一组数据集

	正股关注度	发行规模	主体信用评级	上市价格水平
1	高	50 亿元	AA+	高
2	低	3 亿元	A+	低
3	一般	12 亿元	AA	中
……		……		

数据来源：广发证券发展研究中心

那么，在众多数量化的转债价格分析方法中，为何我们在《转债入门手册》系列中率先选择了机器学习方法进行介绍呢？这是因为，我们曾在《转债入门手册之四——转债定价方法进化史》中提到，由于可转债是一种构成极为复杂的衍生品，简单拆分之下，一只典型的国内市场公募可转债，即是一只信用债和若干美式看涨和看跌期权的集合。在这样的构成下，一方面，基于B-S期权定价思想的主流资产定价方法在国内市场期权定价问题中的适应性存在较大问题；另一方面，各条款的路径依赖特征和相互关联也为转债整体定价增添了难度。

在传统资产定价方法受到限制的情况下，对研究的基本假设更为宽容的机器学习方法成为了新的思路。在传统方法的认知下，机器学习方法在计算过程中较为"黑箱"化，因此在理论和逻辑支撑方面有时会显得缺乏支撑。但对于可转债这样一个复杂的产品，机器学习方法的"黑箱"特性，反而成为了能有效避免方法陷入严苛假定约束的新路径。在国内学术研究方面，Dubrov（2015）和金仁莉（2016）等也已经开启了运用机器学习方法为国内外转债产品进行定价的工作。[3]

方法介绍：随机森林、XGBOOST和支持向量机

MLF超量续作，背后是超储率降至2017年水平MLF超量续作，背后是超储率降至2017年水平MLF超量续作，背后是超储率降至2017年水平

**

在本篇报告中，我们选取了三种常见的学习方法来演示构建机器学习择券体系的思路和流程，分别是随机森林（Random Forest）、XGBoosting和支持向量机。**考虑到本文篇幅有限，各种方法的技术细节也均有大量的公开资料可以参考，因此在本文的方法介绍部分中，我们将着重介绍方法的思想和原理，而更为具体的推导和编程细节，则可参考文中引用的各类参考资料。

**

**

（一）随机森林方法

**

**

随机森林方法是一套在决策树方法基础上，基于集成学习思想发展出的方法体系，因此在介绍随机森林方法之前，我们需要首先对决策树、集成学习方法和Bagging的思想等基础方法有所了解。

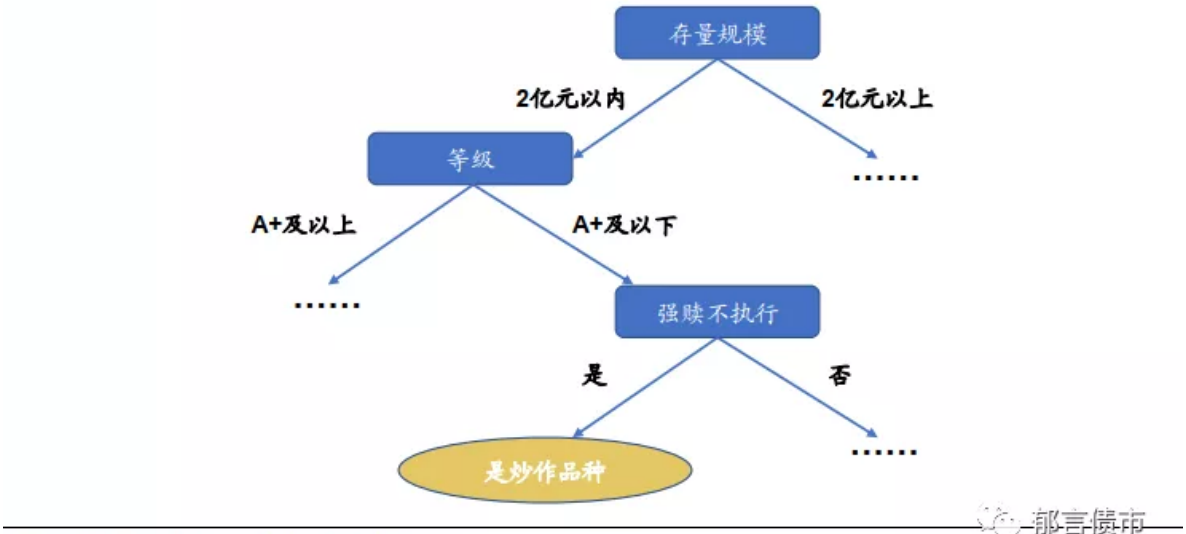
决策树方法，可以理解为一种依次利用研究目标的多个特征属性，来解决其类别划分问题的方法。在该方法中，研究目标的每一个特征会在判断过程中形成一个分化的节点，而这个节点会依照属性表现的不同，派生出新的节点，而在新的节点上，我们又可以依据下一个特征的表现来进行判断，最后形成一个树形结构的判别决策过程。

我们可以用前期转债市场中炒作品种的识别问题为例，来说明决策树的运行原理。假设，我们在判断一只转债是否可能成为炒作品种时，主要参考的指标为：1.存量规模是否在2亿元以下；2.主体信用评级是否在A+或以下；3.是否在强赎条件达成后选择不提前赎回。现在，对于我们的目标个券，我们按照上述的三个关注点分步骤进行测试：我们首先看到，这只转债的存量规模已经小于2亿元；随后，等级方面，我们注意到这只转债的主体信用评级仅为A+；最后，我们观察到这只转债在此前触发强赎条件后并没有发布执行提前赎回的公告，基于上述特征，我们做出判断，这只转债是潜在的炒作品种。而在随后几个交易日，这只转债的确出现了超过1000%的单日换手率，并被交易所认定为异常波动，印证了前期的判断结果。

在上述的决策过程中，每一个节点所测试的内容都包含在上一个节点的范围之内，最终引导出我们对“特定个券是否属于潜在炒作品种”的结论，如图1所示的树形结构起点叫做根节点，而最终分类结果（图1中的黄色圆形节点）为叶节点。那么，如何选择每一个节点对应的属性呢？通常来说，我们在节点的逐层划分过程中，我们希望使得每一个节点分裂后的信息增益（Information Gain）最大，也即是说，使得每一个环节的属性测试能在最大限度上达到标签划分的效果。为了达到这一效果，需要通过数据集来不断优化模型的泛化能力，增强其划分样本外案例的能力，这也是每一个节点最优属性选择的通常原则。具体的节点选择过程中，主要会运用到信息熵（Information Entropy）的概念，本文中不做详细展开。

最后，为了规避过拟合等现象，增强模型对样本外数据的适应性，我们通常还需要对决策树做剪枝（Pruning）等处理。

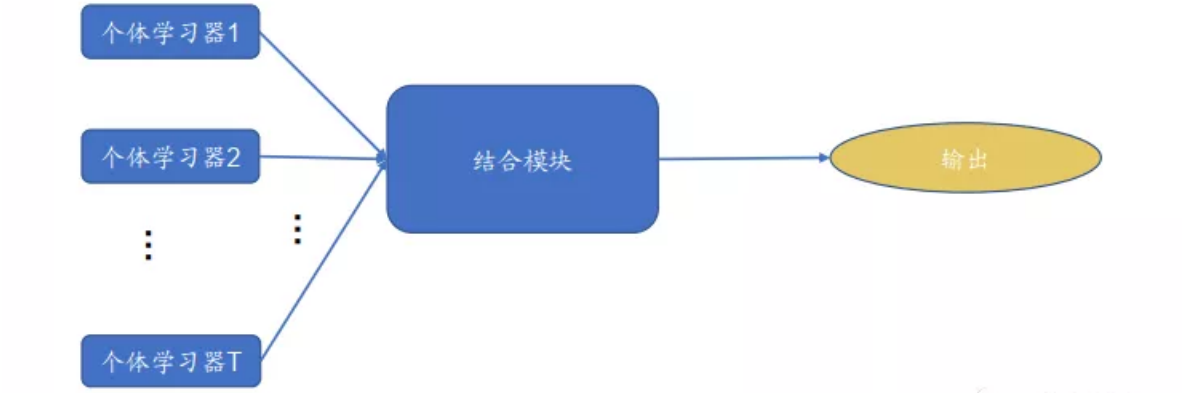
图1：决策树的生成原理



数据来源：广发证券发展研究中心

而**随机森林方法**，则是利用了集成学习（Ensemble Learning）的思想，在决策树的基础上引入了随机属性选择后的一种加强方法。其中，集成学习方法指的是通过建立多个个体学习器，构成结合模块后综合输出决策结果的方法体系（周志华，2016）。在随机森林方法中，个体学习器便是决策树，集合后便会形成一个"决策树集成"。

图2：随机森林的合成原理



数据来源：《机器学习》（周志华，2016），广发证券发展研究中心

随机森林采用的集成方法是Bagging（Breiman，1996a[4]）方法的拓展。Bagging方法源自于自助采样法（BootstrapSampling），其基本思想是：假设我们有若干个数据集，若我们在数据集中重复n次又放回的抽样，就会得到一个n个样本的采样集合（Bootstrap方法）[5]；将以上过程重复T次之后，便可以得到T个含有n个样本的采样集合样本，用每一组样本分别对个体学习器进行训练，再将这些个体学习器进行结合。

具体到随机森林方法上，其在对单个决策树进行集合的过程中，随机森林对Bagging思想做了些许改动，针对决策树的每个节点，首先从所有属性中随机选择k个属性（在炒作券选择的案例中，属性即是规模、等级等要素，这里我们假设属性总共有n），然后从这k个属性中选择一个最优属性用于划分。值得注意的是，当k=d时，其和决策树的构建方法便趋于一致，通常，我们将k设置为log2n（Breiman，

2001a）。

从经验上看，随机森林方法拓展了Bagging方法的多样性，并且在计算过程中的资源耗费程度较小，总体的训练效率通常也优于Bagging，是集成学习中高效方法的代表。

表2中的伪代码展示了随机森林模型在实际运行中的代码编写思路。

表2：随机森林模型的代码编写思路

目的： 生成 C 个分类器

输入： 训练集

过程：

1: ***for*** *i=1 to c* 执行

2: 随机放回地从训练集 D 中抽样，产生 D_i

3: 生成一个对应的根结点 N_i， N_i 包含 D_i

4: 叫做分类器 (N_i)

5: ***end for***

6: 建立树 (N)

7: ***If*** N 仅包含一个分类 ***then***

8: ***return***

9: ***else***

10: 随机从 N 中选择 x %的潜在分割特征

11: 选择具有最高信息增益的分割特征 F 作为分割器， F 有($F_1,...,F_f$)种可能出现的值

12: 将 N 生成 *f* 个子节点， $N_1,...,N_f$

13: ***for*** *i= 1 to f* 执行

14: 使 N_i 包含 D_i， D_i 是 N 中所有匹配到 F_i 的样本

15: 叫做分类器 (N_i)

16: ***end for***

17: ***end if***

输出： C 个分类器

数据来源：广发证券发展研究中心

（二）XGBoosting

Boosting方法与上一节中介绍的随机森林方法一样，同样属于常见的集成学习方法，即核心是多个弱学习器进行联合后，将其增强为一个更强的学习器。Boosting方法的运行原理可以简要概括为：首先利用一套初始的训练集训练出一个效力较弱的基础学习器，再对这套基础学习器的效果进行评估，并对其作出错误判断的样本赋予更高的权重，接着对样本进行调整后再训练出下一个基础学习器，如此往复n次，最后将这n个学习器的判断进行加权结合，得到一个更强的判断效果。

具体来看，为便于理解方法的基本原理，我们选择了经典的AdaBoosting方法作为范例，介绍Boosting方法的运行流程。在AdaBoosting方法中，初始的初始分类器由一组等权的初始样本训练得到。在初始时期，我们假设训练的集样本权重为等权，而在每一次弱学习器得出分类结果后，我们都将依据结果调整下一次训练集中错判样本的权重，使得下一轮学习能够尽可能地修复前一次的错误，并再次开始训练，得到分类结果，调整权重，如此往复。

相较于经典的AdaBoosting方法，我们在本篇报告中实际选用的XGBoosting（eXtremeGradient Boosting，极端梯度提升）方法是一种基于GBDT（Gradient Boosting Decision Tree，梯度提升决策树）的改进方法。原始的GBDT算法基于经验损失函数的负梯度来构造新的决策树，只是在决策树构建完成后再进行剪枝。而XGBoosting在决策树构建阶段就在损失函数中加入了正则项。在介绍XGBoosting的特点介绍中，将不可避免地涉及部分公式推导，只关心方法原理的读者可以将这部分略过。

在XGBoosting中，损失函数被设置为：

其中F(x_i)表示现有t-1的棵树最优解。正则项定义为：

其中T为叶子节点个数，W_j表示第j个叶子节点的预测值。对该损失函数在F_(t-1)处进行二阶泰勒展开可以推导出:

其中T为决策树f_t中叶子节点的个数，

I_j表示所有属于叶子节点j的样本的索引的结合。

假设决策树的结构已知，通过令损失函数相对于W_j的导数为0可以求出在最小化损失函数的情况下各个叶子节点上的预测值：

在本文的回测过程中，我们采用了常用的贪心法来构建出一个次优的树结构，基本思想是从根节点开始，每次对一个叶子节点进行分裂，针对每一种可能分裂根据特定的准则选取最优的分裂。

上述算法流程的编程思路为：

相较于前文中介绍的随机森林方法，Boosting是一种在原理上相对更为简单的集成学习方法。

（三）支持向量机

支持向量机（Support Vector Machine，简称SVM）方法的基本原理，是在样本空间中找到合适的超平面[6]，将具有不同标签的样本尽可能准确地进行划分。我们以一个二维特征的简单划分问题为例，假设在一组数据集中，属性有X1和X2两类（分别为坐标轴的横轴和纵轴），而标记则有“圆圈”和“交叉”两类。显然，在图4的例子中，标记为“圆圈”类别的样本集中在坐标轴的右下角，而“交叉”类别的样本则主

要集中在坐标轴的左上角。

由于上述的划分问题是一个属性维度为二维的问题，即选择了两种特征作为推断依据，因此按照支持向量机的思想，我们需要寻找的划分超平面即是一条2-1=1维的超平面——即一条直线，以尽可能准确地将具有两类标记的样本进行划分。在这里的例子中，其实有很多直线都可以完成这样的划分任务，例如图4中两条蓝色的直线，都可以将两类样本进行完美的划分，但为了尽可能使得这条直线的划分作用对于后续新加入的样本仍然有效，我们应该在众多能完成划分任务的曲线中，寻找处在两类样本最"正中间"的那一条，即图4中红色的直线。这样一来，我们选择的划分超平面将在各条曲线中具有最高的宽容性，对于出现在现有两类样本附近的样本外数据，也将有更高的正确判断机会。

那么，我们要如何衡量划分超平面的"中央"概念呢？我们仍然以前文中的情景为例，在样本空间中，我们可以将划分超平面表达为一个线性方程：

其中，x为属性向量，其中的元素即是前文中提到的X1和X2；w为权重向量，也称为法向量，决定了超平面的方向；b通常意义上的截距项，在这里也叫位移项，决定了超平面与原点之间的距离。

显然，在确定了法向量w以及位移项b之后，超平面在空间中的位置也就随之确定。所以，我们在分类问题中所寻找的，其实就是能使得超平面具有最佳划分效果的法向量和位移项。

若我们找到的划分超平面能将所有的训练样本进行正确分类，那么距离这个超平面最近的几个样本（即图x中标红的样本点）被称为支持向量（Support Vector）。而穿过两类样本的支持向量，并且与超平面平行的虚线与超平面之间的距离，被称为间隔（Margin），为了尽可能地减少意外情况的出现，一个理想的超平面所对应的间隔应该尽可能地宽。这样一来，我们在前文中所述的希望找到处于两类样本"正中间"的划分超平面，也就转化为了寻找得到最大间隔（Maximum Margin）的划分超平面。

由此，超平面的搜寻问题也就转化为了一个约束条件（即使所有类别都能正确划分的w和b）下的间隔最大化问题。若用公式来表示，上述最大化问题即可以由最大化间隔问题转化为最小化w的模长，因为按照欧式空间的距离计算公式，w的模长即是支持向量到超平面距离的分母，而在y为正负1的划分问题中，分子为常数，因此最大化间隔的优化问题也可以表示为最小化w模长的（这也被称作支持向量机的基本型）：

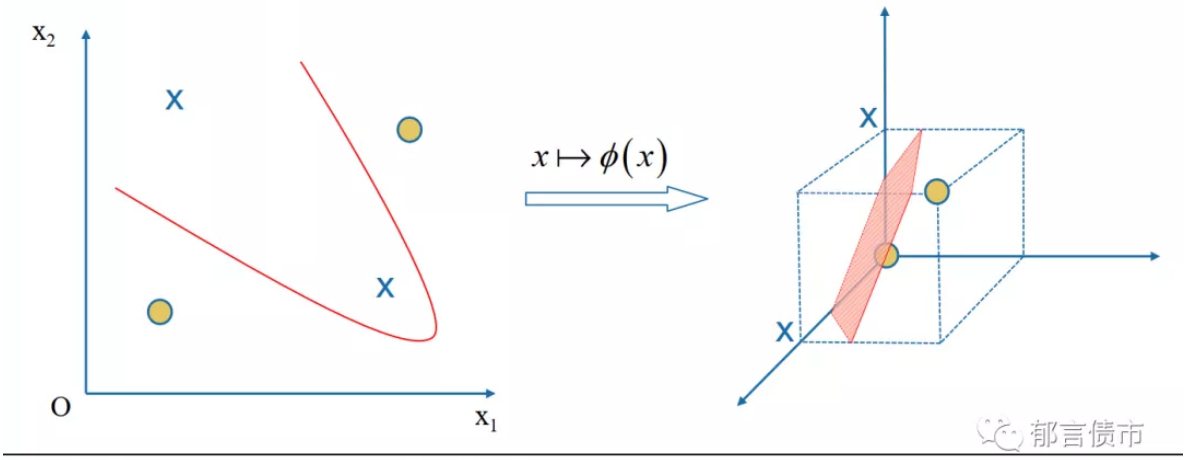
上式是一个凸二次规划问题，在计算中，我们时常会选择实用拉格朗日乘子法得到该规划问题的对偶问题。上述问题的格拉朗日函数可表示为：

其中，α_i为拉格朗日乘子。分别对w和b求偏导之后，即可得到上述问题的对偶问题，之后转化为对的求解：

从前文的例子中，我们用一个二维问题展示了支持向量机的基本原理，但显然，在这里的范例中，划分问题的难度是很低的，因为两类样本的分布都非常集中，我们可以轻易地利用直线进行分割。那么，对于各类样本在空间中相互交叉的场景，我们能找到理想的低维超平面来对样本进行有效划分吗？

这里，我们通常需要将现有二维空间上的样本投射到更高维度的空间，例如在图6所示的例子中，我们在面临一个两类样本无法用直线分隔的问题时，我们便可以将二维样本投射到一个三维的空间上，而在这个三维空间中，我们便可以找到一个适当的3-1=2维超平面对两类样本进行有效分隔，最后再将超平面映射回二维空间，便可以得到一条可以划分两类样本的曲线。事实上，只要划分问题的属性是有限维数的，那么便一定存在一个更高维度的特征空间，使得不同类型的样本可以用n-1维超平面进行划分。

图6：我们可以将样本投射到更高维度的空间，寻找到划分超平面



数据来源：《机器学习》（周志华，2016），广发证券发展研究中心

图6中的Φ(x)表示将x映射到高维空间时的特征向量。在后续的求解过程中，需要涉及到特征向量内积的运算，而当特征空间维数较高时，这一计算过程将变得十分困难，因此计算过程中，时常会需要引入著名的核函数进行计算。后续较为繁杂的计算步骤，本文中暂不讨论，感兴趣的读者可以参考本文的引用书目。

总结来看，支持向量机的思想就是在各类样本空间中，利用各种手段寻找能有效将其划分的超平面，在得到划分超平面之后，对于新的样本，我们便可以通过定位其与划分超平面的相对位置来对其所属的类别进行判断。

支持向量机在具体实现中的编程思路由表2所示（在实际运算中，许多常见语言都会有预先制作好的程序包可以调用）：

表4：支持向量机模型算法

输入： 训练集 X 和相应标签 y，$\alpha \leftarrow 0$

过程：

1：重复

2：**for** all $\{x_i, y_i\}$，$\{x_j, y_j\}$ 执行：

3：优化 α_i 和 α_j

4：**end for**

5：**Until** α 不发生改变或其他限制条件已被满足

输出： 支持向量 $(\alpha_i > 0)$

数据来源：广发证券发展研究中心

回测流程分析

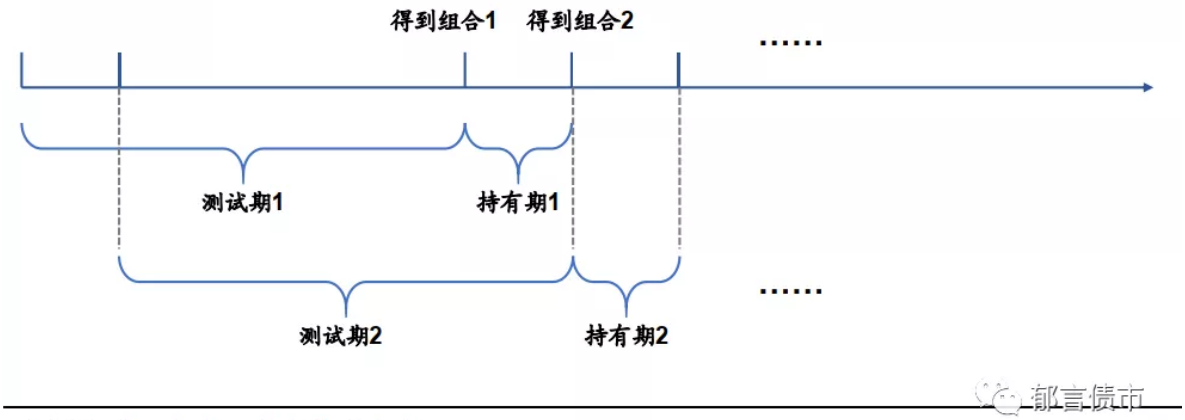
MLF超量续作，背后是超储率降至2017年水平MLF超量续作，背后是超储率降至2017年水平MLF超量续作，背后是超储率降至2017年水平

在本篇报告中，我们分别对前文所述的随机森林、XGBoosting和支持向量机三种方法进行了实际演示，展示了如何利用三种机器学习方法构建一个月度换仓的转债择券策略。

总体来看，本文的测试流程为：数据收集-数据清洗-特征提取-标签构建-拆分训练集和测试集-样本内训练和交叉验证调参-样本外预测-结果分析。

我们构建策略的思路是：首先利用一年的个券面板数据对机器学习模型进行训练，在训练期的末尾用训练出的模型，挑选出最优机会在一个月内取得正向收益的个券等权买入，作为当期的持有组合，并持有该组合一个月。而在持有期的末尾，我们会运用当天前推一个测试器的数据再次对模型进行运算，并在当天得到更新后的持仓组合，如此往复，得到一个月度换仓的滚动回测结果。

图7：滚动回测的思路演示



数据来源：广发证券发展研究中心

在数据方面，我们选择的测试区间为：2017年1月1日至2020年5月15日。

在本文的测试中，我们选用了部分常用的转债行情数据作为转债的属性，分别包括：转债在各交易日的收盘价、纯债价值、转股溢价率、主体信用评级、债券余额以及成交量。[7]

在确定属性数据后，我们按照如下原则对数据进行清洗，以尽可能保证数据质量和个券流动性，并且剔除炒作品种的影响：

删除对应成交量为0的数据；

删除数据长度小于30的债券；

删除A+及以下的债券；

删除存量规模小于1.5亿以下的债券；

删除前20日平均成交量在500万以下的债券。

在标签方面，本文中的标签设置为“过去一个月中，转债价格是否出现上涨”，将过去一个月中价格出现上涨的样本标签标记为1，没有上涨的标记为0。

接下来，我们按照如下流程对训练集和测试集进行拆分：

有效数据集：考虑到随机森林模型、SVM模型对缺失值的敏感性和不稳定性，我们删除所有含有缺失值的数据。加之对于每日可用转债样本量的考虑，最终选取有效数据集为：2018年6月15日至2020年5月15日，共23个月；

训练集：选取2018年6月15日至2019年6月15日，长度为一年的数据作为第一个训练集，之后采用窗口滚动式方法，每次向后移动一个月作为新的训练集，直至2020年4月15日，共计10个训练集；

测试集：只选取每个训练集后的一个交易日作为测试集，共计10个测试集。

在样本内测试环节，我们的处理方法为：分别使用随机森林模型、XGBoost模型、支持向量机模型（SVM）对训练集进行训练。考虑到我们将回测区间按月度划分为10个子区间，因此需要对每个子区间的不同训练集进行重复训练。在每个训练集内（样本内）进行5折交叉验证，选取交叉验证集AUC最好的一组参数作为模型的最优参数，用以预测对应的测试集标签。

在具体的样本外测试方面，我们的处理方法为：确定最优参数后，以对应测试集的特征作为模型的输入，可以得到各只转债超过收益率中位数的预测概率。则我们等量买入预测概率在前十名的转债，持有一个月。在下一个月，通过下一个滚动模型的预测概率，选择换仓。

在回测模型运行完毕之后，我们将综合考虑各个择券模型的累计收益率、AUC（Area under the curve，用于测度模型的预测准确程度）和运算时间等指标综合对模型的效果进行判断。

4

结果分析

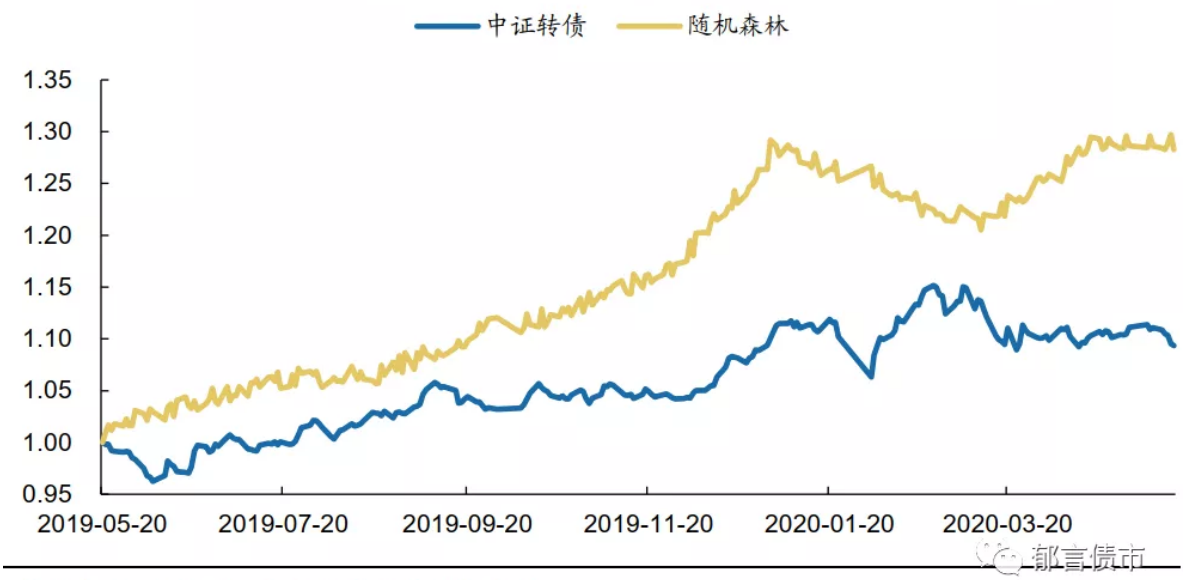
MLF超量续作，背后是超储率降至2017年水平MLF超量续作，背后是超储率降至2017年水平MLF超量续作，背后是超储率降至2017年水平

从累计收益率来看，在样本测试区间中，随机森林方法和XGBoosting方法有着相对较好的表现，其中随机森林方法还在运算效率上具有明显的优势。

首先，从三种策略在回测中的累计收益率来看，随机森林、XGBoosting和支持向量机在样本测试区间均得到了超过中证转债指数的累计收益率，其中随机森林方法的择券表现最为稳定，而XGBoosting方法则在样本观测时期内得到了最高的累计收益率。

具体来看，经随机森林方法筛选的转债组合，自回测开始起累计收益率始终处在中证转债指数累计收益率的上方，即筛选出的组合可以稳定的较中证转债指数取得累计正向收益，测试区间内策略的累计收益率也接近了30%。不过需要注意的是，在2019年末到2020年初，这一策略也曾出现过超过5%的回撤，因此即便是模型回测结果表现相对稳定，仍需注意回撤风险。

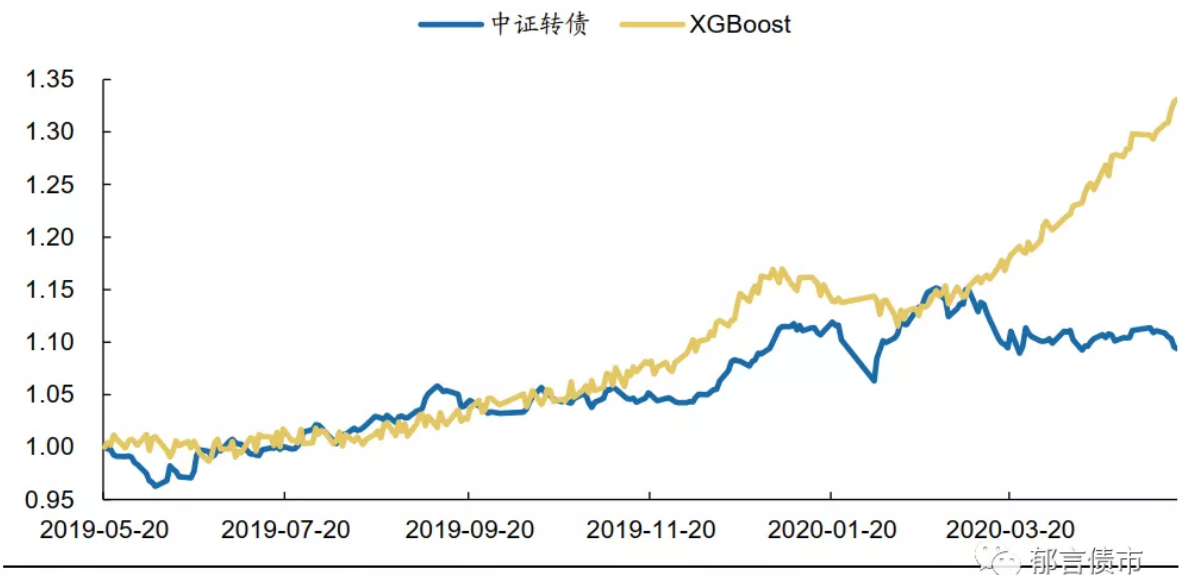
图8：随机森林策略在样本区间内取得了持续超过中证转债指数的累计收益



数据来源：Wind，广发证券发展研究中心

XGBoosting方法构建的组合在样本回测区间取得了超过30%的累计收益率，是三种方法中累计收益最高的。不过需要注意的是，相较于随机森林方法，XGBoosting方法在样本测试区间前段并未取得稳定超过指数的收益水平，回测区间内的高累计收益率主要源于策略在进入2020年后的优异表现。

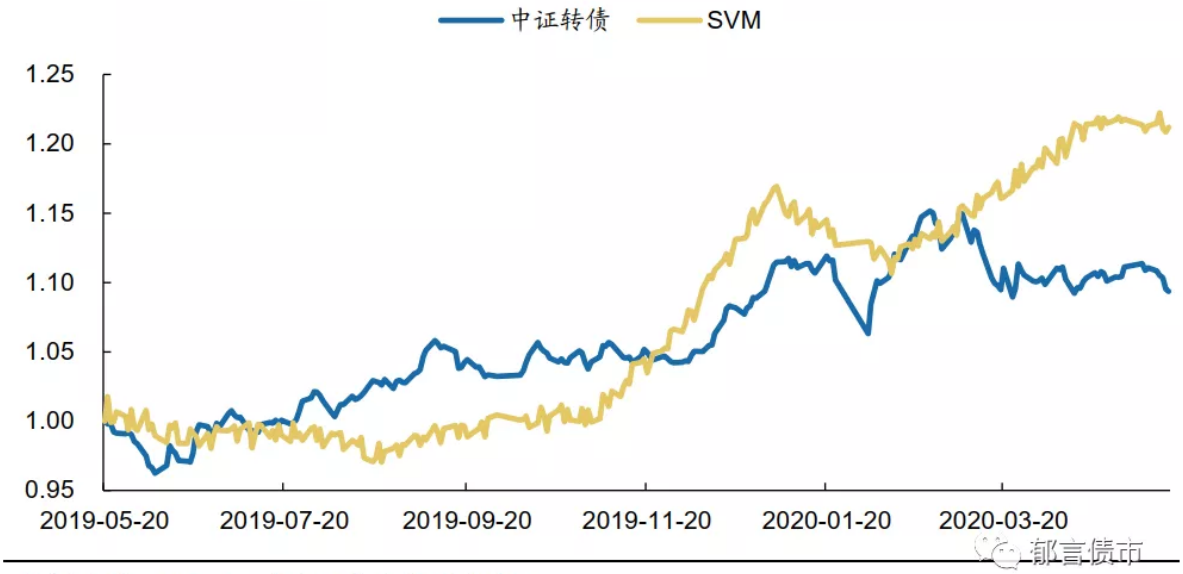
图9：XGBoosting在样本测试区间内取得了最高的累计收益率



数据来源：Wind，广发证券发展研究中心

而SVM方法则在较长时间内并未能够取得胜过指数的收益水平，并且累计收益率也是三种方法中最低的。

图10：SVM择券策略的表现略逊于随机森林和XGBoosting



数据来源：Wind，广发证券发展研究中心

最后，在运行时间方面，随机森林方法具有明显的计算效率优势，同样的回测计算，运用随机森林仅需24.31秒，而XGBoosting[8]和SVM方法则分别耗时38.41秒和34.63秒，虽然总体时间均不长，但随着模型的不断细化，这样的耗时差距会在实际量化投资的过程中带来很大的不同。

而在预测准确度方面，三种方法在样本回测区间内的AUC均处在0.9左右的水平

表5：随机森林和XGBoosting方法的回测结果相对占优

	随机森林	XGBoost	SVM
运行时间	24.31s	38.41s	34.62s
AUC	0.9083	0.9124	0.8991
累计收益率	28.26%	33.09%	21.20%

数据来源：Wind，广发证券发展研究中心

总体来看，在本文的结果演示中，基于机器学习技术构建的择券模型都表现出了超过指数的累计收益，其中随机森林方法在稳定性和运算效率方面占据显著优势，XGBoosting在累计收益水平和判断准确率方面占优，而SVM模型的表现则略逊于随机森林和XGBoosting。但仍需要提醒的是，本文中的模型回测意在方法演示，在量化策略的不断完善中，三种模型的效果也可能随着变量选择的不断完备和运算细节的逐步细化而出现改变。

风险提示：

转债交易和发行相关规则出现重大调整。

**

**

注：

[1]本文的机器学习方法介绍部分参考文献：周志华. 机器学习 [M]. 清华大学出版社, 2016.

[2]假设我们将一年内在Wind发布的券商研报数目>5篇的设定为高，3-5篇设置为中，小于3篇的设置为低。

[3]在参考书目方面，本文的方法介绍部分多以《机器学习》（周志华，2016）为参考资料，部分本文中没有呈现的公式推导和方法细节也可参考原资料。

[4]Breiman，L. (1996). Baggingpredictors. Machine Learning,24(2), 123-140.

[5]当然，这样的采样集合中可能有部分数据集中的样本未被选中，而另一些可能会被选中多次。

[6]超平面指的是n维欧式空间中维度等于（n-1）的线性子空间。在图x的例子中，我们讨论的是一个二维空间中的问题，我们寻找的划分超平面也就是一条一维的直线。

[7]对于主体信用评级这一非数值变量，我们将主体评级按AAA,AA+, AA, A+的顺序分别替换为3、2、1、0。

[8]在实际模型中，可以利用并行等技术手段改进XGBoosting等方法的运算效率。

转债入门手册系列：

《初识转债真面目——转债入门手册之一》

《转债打新全攻略——转债入门手册之二》

《一只转债"老券"的退出之路——转债入门手册之三》

《转债定价方法进化史——转债入门手册之四》

《转债市场参与者行为大盘点——转债入门手册之五》

***已外发报告标题**：*《**如何利用机器学习方法构建转债择券模型——转债入门手册之六》**

对外发布时间：2020年8月24日

报告作者：

刘 郁，SAC 执证号：S0260520010001，SFC CE No.BPM217，邮箱：shliuyu@gf.com.cn

联系人：

田乐蒙，邮箱：tianlemeng@gf.com.cn

广发固收刘郁团队

法律声明

请向下滑动参见广发证券股份有限公司有关微信公众平台推送内容的完整法律声明：

阅读 4298

赞47在看54