基于分类器模型的选股策略研究

百瑞观点：分类器模型是一种基于统计学习理论的模式识别方法，在量化投资领域的应用中可以筛选有效的因子，实现对股价走势进行预测并构造个股组合，因此，可以将分类器模型理解为多因子模型的另一种策略构造方法。多因子选股模型长期以来都是量化投资者共同关注的领域，无论是在因子构造方面还是在利用因子预测收益的方法选择方面，量化研究工作者都投入了大量的精力。在因子选择方面，因子的数量和种类在经过长期的发展之后已经逐渐趋于标准化，在商业化的因子数据提供商不断涌现的大趋势下，量化研究更倾向于将重点放在策略构造方法的研究中。

资本市场随着投资者的不断涌入以及金融科技的不断发展，交易时所涉及到的信息也逐渐繁杂。因此，各个领域的成功经验都被不断地拿到量化投资领域来进行实践，人工智能，机器学习因为他独特的优势也被量化投资研究者所青睐。机器学习的核心在于其使用的模型和目标函数，而分类器模型因为具备丰富的算法，被广泛的应用到了机器学习中，实现了机器学习在各个领域的应用。在量化投资领域中使用以分类器模型为核心的机器学习算法可以将众多因子有效的组合到一起，发挥各自的优势，从而达到选择出在统计上具有超额收益的个股组合。

一、分类器模型概述

（一）分类器的原理

分类器是数据挖掘的一种重要方法，分类器的概念是在已有数据的基础上学会一个分类函数或构造出一个分类模型，该函数或模型能够把数据库中的数据纪录映射到给定类别中的某一个，从而可以应用于数据预测。

分类器的分析的数据是特征值，通俗点而言就像在基本面分析中判定哪些是未来影响产品价格的因素，比如季节因素、下游产品价格等，这些因素除了其自身属性的不同外还应该具备在同级影响因素中具备较低的相关性，也就是最终分类的因素都会对价格进行影响，但是彼此之间的影响相对较小。而预测则是根据特征值对已知分类对象进行分析，从而对特征值相似的未知分类对象的类别进行预测。

作为机器学习的核心，分类器模型的程序目标是使得模型在执行的时候提升它在某项任务上的能力，而非直接编写程序的固定行为。分类器模型包括多种问题的定义，根据不同领域的问题提供很多种不同的算法来解决。

（二）分类器的应用

分类器模型的应用覆盖了现在机器学习模型应用的大部分领域，包括图像识别，关键词推荐，移动设备用户的行为识别（跑步、上楼、

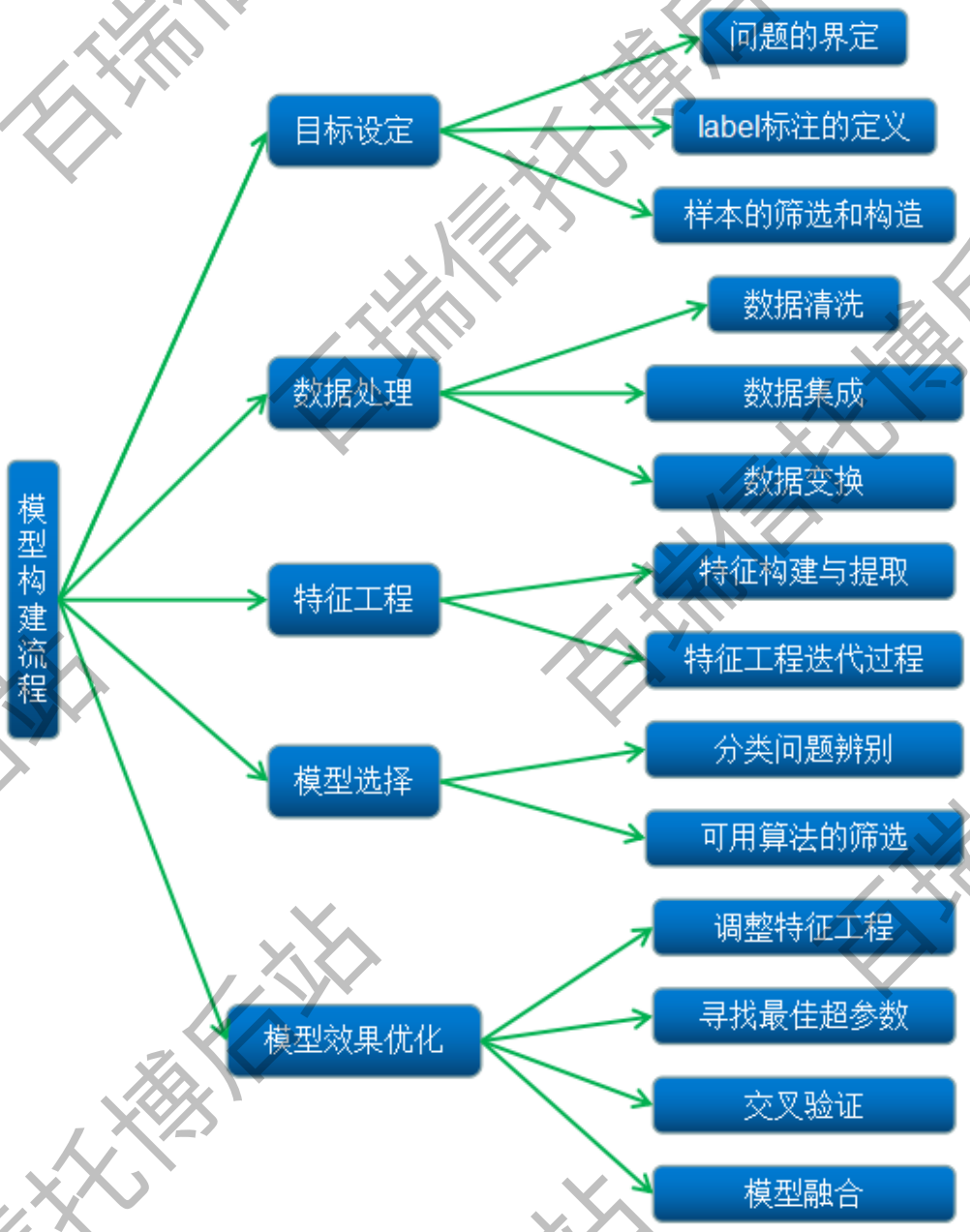
下楼等行为的识别）。诸多应用所使用的分类器模型也有所不同，选择合适的分类器主要考虑的是应用场景。2017 年以来，一些传统的多因子选股模型遭遇了较大回撤，其中市值、反转等因子风格转变的显著性和持续时间均超乎预期。因此，如何预测因子的有效性并进行因子权重配置的调整成为投资者关心的课题。传统的量化模型使用的是线性回归分析，线性回归分析能够明确地解答“多少”的问题，可以解答输出为连续数值 的问题。当预测的输出是离散的 类别时，这个监督学习任务就叫做分类。例如我们可以把上市公司数据抽取出若干特征，如市值，销售净利润，股票价格的各种指标以及是否有负面报道，利用分类预测下个阶段该公司的股票收益率是否能胜出其他上市公司。

目前这种量化建模的方式也已经在众多私募、券商等中应用已久，但是对于大多数个体投资者而言，还是一个十分陌生的领域。因此，文章将展开这个建模方式的过程，介绍如何建立分类与预测模型，辅助投资者对自身分析逻辑中的分析框架进行量化分析，方便其多元化的交易分析。

二、构建分类器模型的流程

机器学习分类器模型的原理简单易懂，构建模型的流程也简单明了机器学习的建模核心流程主要包含目标设定，数据处理，特征工程，模型选择以及模型效果优化（如图一所示）。一个模型能否有效预测的重点在于每个环节如何设定并调整模型实现最优效果。如果将分类

任务比做纺织，目标设定就相当于为我们最终做出的布设定一个标准，数据处理过程相当于为这个标准选择合适的材料，特征工程就相当于将这些材料编制成纺线，模型选择就是为我们的纺织目标选择纺织机，最后的模型效果优化就是将这一过程中的每一个环节有效的衔接起来。接下来介绍一下完整的分类器建模流程。



图一：分类器建模流程

三、目标的设定

组建机器学习分类器模型时首先要确定需要这个模型来解决什么问题，以投资为目标进行建模也需要先将投资方法细化，并为细化后的结果标注进行定义。以投资股票为例，目标需要设定预测的入场时间和出场时间，确定持股多久为一个交易周期，随后，需要定义一个交易周期中实现的不同交易水平如何分类。例如，对于中证 500 对冲模型，我们设定 5 个交易日为一个交易周期，样本在一个交易周期的收益高于中证 500 股指期货的同期收益为“1”类，其他为“0”类作为分类标准。

确定研究问题并定义好分类的标注后，需要进行样本的筛选和构造，在这个阶段，会确定我们可用的数据样本的规模，确定训练样本和测试样本拆分的方式，在投资模型的历史回测过程中，训练样本和测试样本是以滚动窗口（rolling window）的方式持续更新的，即每天将前一天的测试样本放入总的训练样本中，并取出新一天的数据做测试样本。在这个构架中的数据是可以用来判断股票走势的因子，包括基本面因子，技术因子，舆情因子等。

四、数据处理

（一）数据清洗

数据清洗是指发现并纠正数据可识别的错误的最后一道程序，需要处理数据中存在的无效值、缺失值以及噪声值。其中，噪声数据是

指一个测量变量中的随机错误和偏差。通过数据清洗，可以使纳入模型的数据在准确率、完整性、一致性、时效性、可信性和可解释性上满足应用的要求。使用机器学习方法进行模型设计时，不一定要摒弃所有传统计量模型的方法，尤其是在数据清洗的过程中。虽然流程繁琐，但是相较模型调参和预测来讲略为简单，传统计量模型的算法可以使得数据处理过程更加易于分析，在对于单一因子的处理和因子效果的分析比较中具备一定的优势。

1.缺失值的处理

缺失值的处理，主要方法包括删除法和插补法。删除法中包括删除样本，删除变量以及改变权重等方法。插补法常见的有均值插补、回归插补、二阶插补、热平台、冷平台等单一变量插补法。其中，均值法是通过计算缺失值所在变量所有非缺失观测值的均值，使用均值来代替缺失值的插补方法。均值法不能利用相关变量信息，因此会存在一定偏差，而回归插补是将需要插补变量作为因变量，其他相关变量作为自变量，通过建立回归模型预测出因变量的值对缺失变量进行插补。热平台插补是指在非缺失数据集中找到一个与缺失值所在样本相似的样本（匹配样本），利用其中的观测值对缺失值进行插补。在实际操作中，尤其当变量数量很多时，通常很难找到与需要插补样本完全相同的样本，此时可以按照某些变量将数据分层，在层中对缺失值使用均值插补，即采取冷平台插补法。

2.噪声数据的处理

噪声数据对于某些分类模型的分析结果影响很大，例如聚类模型和逻辑模型，但是对决策树、神经网络以及支持向量机的影响较小。虽然通过结合多层神经网络形成的逻辑模型和聚类模型受异常值对最终分析的影响也比较小，但是异常值的存在可能会使一些因子失效。另外，对于特定的分类器，噪声数据处理一方面可以使原始数据中不存在错误值及偏离期望的孤立点至，另一方面可以使最终形成的特征值落于模型判断最敏感的分布区域。

在处理噪声数据之前先要进行数据检查，之后才能根据噪声数据的特征进行处理。数据检查的方法分为两种，第一种方法是使用传统多因子模型的方法，即对符合正态分布的数据标准化小概率数据，对于偏度较大的数据可以选择用$\log_1 x$函数进行转化，使其更加服从正态分布，或者使用因子载荷原始值对数据进行标准化处理。第二种检测方法是分类模型中的聚类算法，通过机器学习的训练将类似的取值组织成“群”或者“簇”，落在“簇”集合之外的值则被视为离群点。在进行噪声数据检查之后，需要根据噪声分布的特点进行数据处理。

以股票价格的噪声处理为例，由于股票市场动态的复杂性，股票价格往往充满大量的噪声数据。这对分类模型的学习效果会产生一定的影响，因此我们需要将噪音的影响从股票价格的趋势和结构中剔除出去。首先，股票回报（return）的分布偏度较大，使用$\log_1$return 处理回报数据，可以使其趋于标准分布，之后可以进行降噪处理。由于股票回报值分布的非平稳性以及需要含有局部区域的时间信息，我们选择小波变换进行数据降噪处理：

$$R_{(a,b)}=\frac{1}{\sqrt{a}}\int_{-\infty}^{\infty}r(t)\times\psi\left(\frac{t-b}{a}\right)dt$$

其中 a 是缩放尺度，控制小波函数的伸缩，b 是平移参数，控制小波函数的平移量，$\psi\left(\frac{t-b}{a}\right)$是由小波母函数$\psi(t)$经过以 a 为伸缩因子和 b 为平移因子进行变换后得到的一个函数簇，小波母函数的基本特性为$\int_{-\infty}^{\infty}\psi(t)dt=0$，$r(t)$是原回报值，$R_{(a,b)}$是以 a 为尺度因子、b 为平移因子进行降噪处理后的回报值。从形式上也可以看出，原回报值得小波变换本质上是原来的$r(t)$在 t=b 附近按$\Psi_{a,b}(t)$进行加权平均，体现的是以$\Psi_{a,b}(t)$为标准$r(t)$快慢变化的情况。

（二）数据集成与数据变换

将多个数据源中的数据结合起来并统一存储，建立数据仓库的过程实际上就是数据集成。对不同来源的数据进行集成时必须注意数据结构的调整，将不同来源、格式的数据进行统一处理才能应用于模型。

数据变换包括两个步骤，第一部是数据算法扩张。对分类器模型而言，我们在使用多因子数据进行分析时每一个因子都是独立存在的，分类器模型并不能找出因子之间的关联。因此，我们需要找出相关的因子并构造出他们的关联因子。例如多因子模型中一部分因子为估值而设计，比如经营性现金流（OCFP）的增长率/自由现金流（FCFP）增长率的值作为因子，可以跟踪现金流的性质变化。这一过程可以通过分类器的聚类算法进行自动变换，将具备类似属性的值组织成“群落”，之后在“群落”中进行多种交叉运算形成关联因子。

第二部是数据规约，是指在尽可能保持数据原貌的前提下，最大限度地精简数据量。通常用维归约、数值归约方法实现。由于我们在算法扩张中形成的因字数可以高达几万个，这些因子之间可能存在高相关性，而且大量的数据直接放入分类器模型中会增加运算的时间，因此需要通过维规约移除相关性高的因子从而提高模型效率。常见的维归约方法有：高相关滤波、主成分分析法、映射方法、神经网络和聚类方法等。在初步的数据处理中，根据多因子模型的数据的特点，我们选择高相关滤波可以最快的速度去除共线性高的数据。高相关滤波认为当两个因子变化趋势相似时，他们包含的信息也相似。这样使在相似因子中选择其中一列就可以满足机器学习模型的训练需要。我们只需要建立因子库的相关系数矩阵，在相关度高的因子中进行二选一剔除即可。

五、特征工程

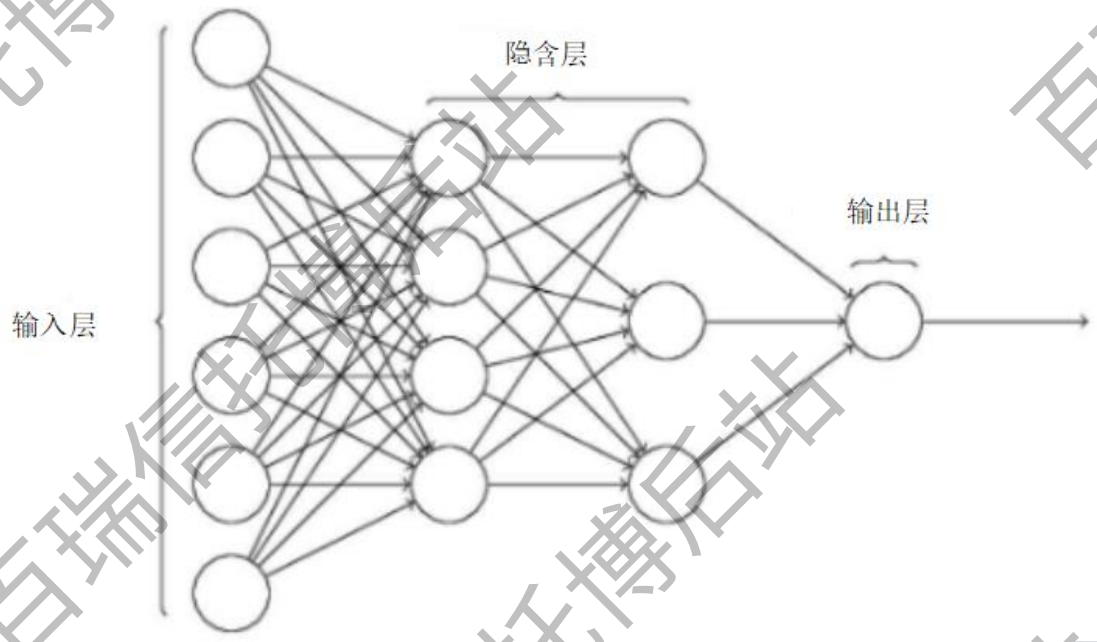
特征工程是将原始数据转化为特征，更好的表示预测模型处理的实际问题。我们数据库中的各种因子只是一列参数，我们可以根据研判标准来设计它的特征值，从而形成带入分类器模型中的真实因子。数据特征会直接影响预测模型的选择和实现的预测结果，大多数模型都可以通过优质的特征得到很好的学习结果。为后面选择模型和调整最优参数降低难度。

（一）特征值得的构建与提取

数据处理过程中我们已经形成了干净的因子库。一部分因子与股票价值评估存在直接关系，比如估值因子和成长因子，除了将因子本身作为特征值之外，只需要将因子的变化率等趋势性的特征进行构建并加入特征库。而另一部分因子，需要根据因子的特性为股票走势预测专门提取特征值，比如技术指标因子。以超买超卖型指标为例，CCI,KDJ 因子走向本身对股票价格的预测性并不高，需要根据因子的走势提取超买超卖信号作为特征值带入模型。

（二）特征工程迭代过程

这一过程是机器学习大量减少特征处理的重要步骤，从功能上来讲特征工程迭代是神经网络模型构建的滚动窗口降维方法。我们最终做出的特征值库变量非常多，直接带入分类器模型中计算量仍然很大。另外，对于股票预测来讲，不是所有特征都能在所有时间都是有效的，那么我们就需要通过特征工程迭代过程来为每一个交易日选择有效性最强的特征值进行选股。特征工程迭代的方法有很多种，考虑到多因子选股策略的数据特征，我们使用多层神经网络模型进行特征迭代。



图二:多层神经网络结构

如图二所示，多层神经网络模型进行每一次模型决策之前，先将所有特征值都带入输入层，在每一层训练时都随机带入比前一层少一定数量的特征值，对每个特征值进行多次的迭代测试来确定最有效的特征值组。通过多个隐含层对特征值组进行多次的训练，筛选出最优特征组合得以使分类器模型做出最后的输出层判断。在这一过程中，分类器模型需要在每一层都参与选择才能够筛选出有效因子。通过历史迭代的过程，我们可以统计被采用频率最高的特征值，并可以依此分析不同市场环境中有效特征值的分布情况。接下来我们需要选择合适的分类器模型。

六、分类器模型的选择

没有最好的分类器，只有最合适的分类器。在分类器选择过程中需要根据数据情况和分类器的特性来选择合适的分类器，目前较为广

泛应用的分类器有 KNN，贝叶斯，决策树，随机森，SVM 分类器和 SOFTMAX 等分类器，基于我们的分类问题和数据环境可以对分类器进行筛选。

（一）分类器的考量标准

在多因子选股模型选择分类器的时候主要考虑三个要点：泛化能力和拟合之间的权衡，分类函数的复杂度和训练数据的大小，输入特征空间的维数以及输入的特征向量之间的均一性和相互之间的关系。

1.泛化能力和拟合之间的权衡

过拟合评估的是分类器在训练样本上的性能。如果一个分类器在训练样本上的正确率很高，说明分类器能够很好地拟合训练样本。但是一个很好的拟合训练样本的分类器就存在着较大的偏置，所以在测试样本上不一定能够得到好的效果。如果一个分类器在训练样本上能够得到很好效果但是在测试样本上效果下降严重，说明分类器对训练样本拟合过度。若分类器在测试样本上能够取得好效果，那么说明分类器的泛化能力强，这是选股策略更重视的能力。分类器的泛化和拟合是一个此消彼长的过程，所以分类器需要在泛化能力和拟合能力间取得平衡。

2. 分类函数的复杂度和训练样本的大小

训练样本的大小对于分类器的选择也是至关重要的，如果是一个简单的分类问题，那么拟合能力强泛化能力弱的分类器就可以通过很小的一部分训练数据来得到。反之，如果是一个复杂的分类问题，那么分类器学习就需要大量的训练数据和泛化能力强的学习算法。一个好的分类器应该能够根据问题的复杂度和训练数据的大小适当地调整拟合能力和泛化能力之间的平衡。

3. 输入的特征空间的维数

如果输入特征空间的向量维数很高的话，就会造成分类问题变得复杂，即使最后的分类函数仅仅就靠几个特征来决定的。这是因为过高的特征维数会混淆学习算法并且导致分类器的泛化能力过强，而泛化能力过强会使得分类器变化太大，性能下降。因此，针对选股模型的数据特点，特征空间的向量维数并不低，相对简单的分类器模型难以实现兼顾拟合能力和泛化能力。需要考虑适应能力较强的相对复杂的分类器。另外，通过复合多层神经网络模型可以减少高维数据对分类器模型的影响。

（二）可用分类器的筛选

1. KNN 算法

KNN 邻近算法是数据挖掘分类及湖中最简单的算法之一，所谓邻近，就是说每个样本都可以用它最接近的 k 个邻居来代表，这个方法在分类决策时，只与及少量的相邻样本有关，对异常值敏感，容易受

到数据不平衡的影响。对于股票长期历史的大样本数据来讲，K 临近方法寻找附近样本有其局限性，难以实现我们的分类目标。由于 KNN 方法主要靠周围有限的邻近的样本，而不是靠判别类域的方法来确定所属类别的，因此对于类域的交叉或重叠较多的待分样本集来说，KNN 方法较其他方法更为适合。

2. 贝叶斯分类器

贝叶斯分类器的设计方法是一种最基本的统计分类方法，其分类原理是通过某对象的先验概率，利用贝叶斯公式计算出其后验概率，即该对象属于某一类的概率，选择具有最大后验概率的类作为该对象所属的类。适用于不同维度之间相关性较小的时候，比较容易解释。也适合增量训练，不必要再重算一遍，比如垃圾邮件处理的应用。而贝叶斯算法在相关性较强的样本及维度上训练效果很差，由于股票数据不同纬度之间相关性会很大，进行每日因子重新权重时处理因子相关性的工作量也很大，所以使用贝叶斯分类器并不在我们的考虑范围里

3. 决策树、随机森模型

决策树模型的特点在于沿着特征做切分，随着层层递进，这个划分会越来越细，举个简单的例子，当我们预测一个股票的走势时，决策树的第一层可能是这个上市公司的市值，属于大市值就走左边的的枝权进行进一步预测，小市值则走右边的枝权，这就说明市值对股票走势有很强的影响，随机森就是决策树的集成算法，有很多成功的量

化私募公司将随机森作为多头选股模型的组成部分。它首先随机选取不同的特征和训练样本生成大量的决策树，然后根据这些决策树的结果来进行最终的分类。这一模型的数据维度相对比较低（几十维），而且基本上不需要很多参数调整就可以达到不错的分类效果。这要求对股票因子挖掘的非常细致，每次进行预测时最终带入不超过一百个因子为佳，我们的大数据挖掘可能会形成几千个因子，在随机森模型中训练速度受限，所以需要寻找更高效的分类器。

4.SVM 与 SOFTMAX 模型

SVM 和 SOFTMAX 分类器都起源于逻辑模型，其中 SVM 分类器也叫支持向量机，目前被很多量化私募公司作为主要模决策型，SVM 的核心思想就是找到不同类别之间的分界面，这样可以将两类样本尽量落在面的两边且与界面距离最远。SVM 是通过支撑面做分类的，也就是说不需要计算所有的样本，高维数据中只需去少量的样本，节省了内存，模型可以解决高维问题和非线性问题，在舆情因子挖掘中尤其受欢迎。SOFTMAX 分类器解决的主要问题多分类问题，即我们可以将股票将来的走势分为多个类别，用涨跌幅做分类标准为例，就是下一个持有期间这只股票从跌 10%到涨 10%可以分成多个类别，对这只股票的走势进行更细分的分类预测。对分类的调整可以依据最终带入模型的数据进行设置，从而获取你和能力和泛化能力之前的平衡。SVM 在很多数据集上都有优秀的表现，在文本挖掘中有更强的实用性。而

SOFTMAX 模型胜在模型清晰，背后的概率学经得住推敲。它拟合出来的参数就代表了每一个因子对结果的影响。

这两个模型可以从拟合能力上做进一步筛选，分类器模型的拟合过程依靠自身的损失函数，SVM 和 SOFTMAX 的损失函数在做损失值评估时有所不同，SVM 的训练目的是满足边际，而 SOFTMAX 则是永不满足。比如在做车辆类型区分时，轿车图片和货车图片会产生相近的频分，它就会学习轿车与货车之间的区别。如果给它一个青蛙的图片（青蛙属于被标注为‘其他’的分类），那么‘其他’类的分数只要比车辆类大过某个边际值，模型就不会学到什么东西。也就是说 SVM 只会从评分相近的目标样本中进行学习。这在语言分析中比较有优势，但是选股模型中样本特征可以相差很大，我们需要同时兼顾评分差距较大的样本训练。SOFTMAX 分类器来讲，即使“其他”类的分数已经远高于车辆类很多，它依旧会尝试增加这个计算差距，在对股票进行分析时，能对表现特别优异或者特别差的股票进行更准确的评估。因此，在可用分类器模型中，SOFTMAX 是首选的分类器模型。

七、模型效果优化

（一）调整特征工程

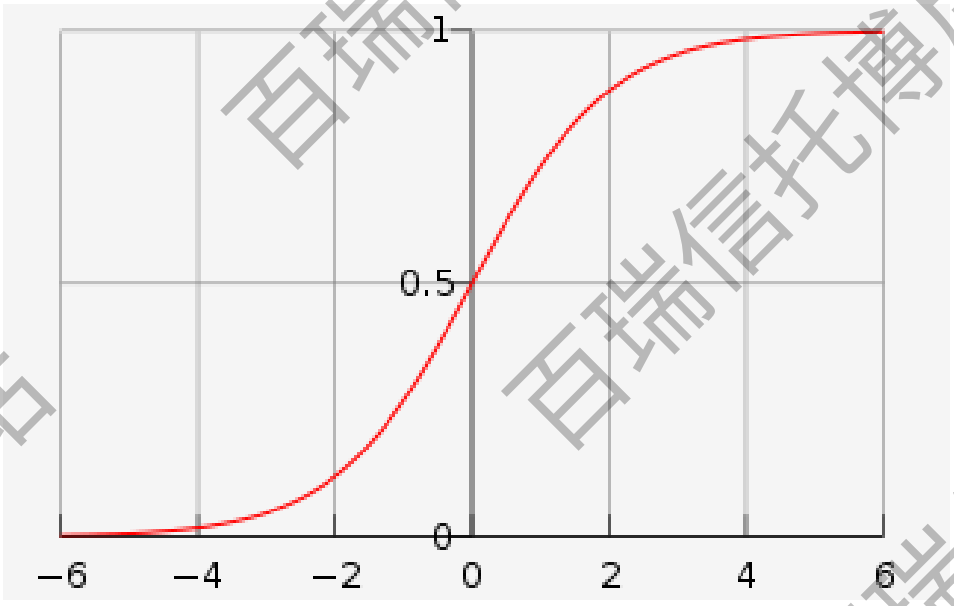
分类器模型的选择看似复杂，但是掌握大致的方向即可选出合适的分类器，而特征工程的调整实则是更加重要的步骤。我们需要将之

前构建出的特征值通过一定的算法形成新的数据，使所有的数据位于所选分类器辨别最为敏感的数据区间。

对于我们所选出的 SOFTMAX 模型而言，它具有一个激活函数的核心部分，所有的拟合训练以及参数调整都围绕着这个核心展开：

$$h_\theta(x)=\frac{1}{1+e^{-\theta^T x}}$$

其中 x 是 n 维特征向量，函数 g 就是 logistic 函数。这个核心函数的图像如下图所示：



图三:Logistic 函数

简单来讲，这个函数就是用我们所给出的特征值组合（x）来对分类（0 或 1）做以辨别，我们可以看到，在 x 大于 4 或者小于-4 的范围中，模型的辨识速度开始锐减，在 0 附近的范围中对分类倾向于

0 还是 1 有更敏感的变化速度。因此，为了使模型能够更清晰的辨别因子对股票分类的影响，我们需要根据每一个因子的特征，将其总体分布投射在更小的范围中，投射的方法需要根据因子的特征单独设计。

（二）寻找最佳超参数

当数据和模型都调整好之后，我们可以先进行模型试跑，在试跑的过程中找出存在的问题。针对我们所选出的 SOFTMAX 模型而言，我们可以选择调整的超参数主要有学习率、学习率衰减速度、正则系数，对于复合的神经网络模型，我们需要调整的超参数主要有隐含层的神经元个数，最小训练样本和神经网络层数。

SOFTMAX 模型的学习率的调整是为了更快更好的降低损失值，即降低每一次选错的概率。学习率的大小代表每次学习对普通参数调整的变化，学习率越大每一次调整的越多，或许会更快的降低损失值，但也很有可能会在最优解左右徘徊。这时，可以设置学习率衰减速度，根据损失值的变化情况可以让步伐变小并逐步靠近最优解。

这一过程逐渐靠近拟合度的最优解，每次训练后我们都需要在测试样本中验证是否准确率相近，如果训练样本准确率一直大幅高于测试样本，则说明发生过拟合情况。过拟合在训练数据不够多或者训练次数过多的时候都会出现，这时就需要正则化的惩罚参数来调整训练参数，保持训练集准确率和验证机准确率相近。这相当于保证我们的

模型在历史模拟的成功率要与投资实际的成功率不相上下。正则化有两个方面改动，一方面使用权重衰减的方法来调整分类器模型的参数。过拟合，就是拟合函数需要顾忌每一个点，最终形成的拟合函数波动很大，在某些很小的区间里，函数值的变化很剧烈。而正则化是通过约束参数的范数使其不要太大，所以可以在一定程度上减少过拟合情况。

正则化的过程另一方面是直接修改神经网络本身。我们称作 Dropout 过程，就是在每次训练时随机删除一部分训练样本。使用这个过程相当于训练了很多个只有部分隐层单元的神经网络，每一个这样的网络，都可以给出一个分类结果，这些结果有的是正确的，有的是错误的。随着训练的进行，大部分网络都可以给出正确的分类结果，那么少数的错误分类结果就不会对最终结果造成大的影响，这个过程产生在输入层和隐含层之间。

对于神经网络模型的参数调整，还包含隐含层的神经元个数，最小训练样本和神经网络层数。神经网络模型的优势在于每一个交易周期都可以重新对因子进行训练赋权，每一个周期市场环境在产生变化，因子的有效性也会随之变化，甚至有些因子对股票预测的相关性会由正转负。在隐含层中我们会进一步缩减有效因子的选择，挑选有效性最强、最稳定的因子。那么就要对隐含层的数量以及每一层加入训练的神经元个数进行调整，并且依据训练效果调整最小训练样本以增加整体模型的训练速度。

（三）交叉验证和模型融合

交叉验证和模型融合对模型最终有效性做完整的考量，在超参数的调整过程中，我们需要对整个模型进行交叉验证并得出最优解。另外，根据投资策略的目的，还可以在模型的输出层融合更多的模型来实现更好的分类结果。

二分类模型的输出部分，我们可以得到四种结果，真正类，假正类，真负类和假负类，可以根据分类结果对模型进行优化。从多头选股策略的角度来看真正类率是主要的优化目标，真正类率是真正类在正类中的比例，即预判为涨的股票中选对的概率。这个概率高了，即便选出的股票数量很少，仍然能保持着高胜率在做投资。以优化这个概率为目的对模型进行整，可以在每次训练中抓取分类为正类中的假正例样本，对假正例样本的因子效果进行分析，弱化给予假正例样本高评分的因子。可以对正例样本叠加一个分类器模型进行优化，并在假例样本中进行交叉验证，检测叠加模型的有效性，也可以在整体样本中用优化后的因子权重进行重新分析。多分类模型的输出，可以看做 n 个二分类器输出的矩阵，对于多分类器的优化，需要根据策略的设计具体调整。

除了融合更多复杂的模型的方法之外，还可以用模型本身的特征进行优化。通过 SOFTMAX 分类器模型进行分析，可以得到模型对每一个分类的评分。因此，我们还可以在机器学习模型尾部添加一个提高真正类率为目标的算法，对每一部分做模型验证最终得出有效的投资

策略模型。基金管理者需要根据投投资策略具体设计分类方法，并根据分类的效果对模型做更细节的调整。

八、应用延伸

分类器的应用非常广泛，其实它在金融领域中的应用也并不止于多因子选股模型。在智能投顾系统中，我们可以用分类器模型做客户画像。用分类器模型对客户的风险承受水平、收益目标以及风格偏好等要求进行分析，对客户类型进行区分并自动推荐最适合的金融产品或者投资组合。在消费金融的风控系统中，我们也可以用分类器模型对客户的收入、消费习惯等信用水平相关的信息进行分析，对客户的风险进行分类并针对不同的风险水平匹配相应的限制条件。