一种基于高阶奇异分解的个性化股票推荐算法

茅斯佳[1,2] 臧斌宇[1,2] 张 谧[1,3]

[1](复旦大学软件学院 上海 200433)

[2](复旦大学软件学院并行处理研究所 上海 200433)

[3](上海市智能信息处理重点实验室(复旦大学) 上海 200433)

摘 要 提出算法预测基金经理对股票的投资策略,为个体投资者提供投资意见。不同于仅依据股票本身信息推荐的传统算法,该算法通过高阶奇异值分解算法 HOSVD(Higher Order Singular Value Decomposition)学习基金经理的历史交易记录和投资者的个人特征因素,为投资者提供个性化推荐。除此之外,将非个性化推荐与个性化推荐进行整合,进一步提高推荐质量。对真实股票交易数据的仿真实验结果表明,用于推荐的个性化算法在准确度和收益率方面,优于传统的非个性化算法。

关键词 股票推荐 高阶奇异分解 线性回归

中图分类号 TP3 **文献标识码** A DOI: 10.3969/j.issn.1000-386x.2015.10.068

A HOSVD-BASED PERSONALISED STOCK RECOMMENDATION ALGORITHM

Mao Sijia[1,2] Zang Binyu[1,2] Zhang Mi[1,2]

[1](*School of Software, Fudan University, Shanghai 200433, China*)

[2](*Parallel Processing Institution, School of Software, Fudan University, Shanghai 200433, China*)

[3](*Shanghai Key Laboratory of Intelligent Information Processing, Fudan University, Shanghai 200433, China*)

Abstract Through predicting fund managers' investments strategy on stocks, the algorithm helps the individual investors in making rational investments decisions. Unlike traditional algorithms that solely based on stocks' information, the algorithm learns from historical transactions record of fund managers as well as the factors of personal features of investors through high-order SVD (HOSVD) algorithm to provide the personalised recommendation for investors. Besides, for further improving recommendation quality, it integrates the non-personalised and personalised recommendations. Results of simulation experiment on a real-life stock transaction dataset show that compared with traditional non-personalised algorithm, the personalised algorithm used for recommendation gains a better performance in precision and yield rate.

Keywords Stock recommendation High-order singular value decomposition (SVD) Linear regression

0 引 言

随着我国金融市场尤其是股票市场的迅猛发展,越来越多的个人和机构开始进入股票市场进行价值投资,而股票市场的波动也随之变得更为剧烈,存在较大的风险,由此对证券投资的研究也应运而生。人们尝试通过机器学习以及数据挖掘等一系列方法对股价的走势进行预期判断,从而达到预期的投资组合收益。

如今金融界中投资机构主要基于金融时间序列模型对投资组合进行建模,同时针对不同投资人的自身特点进行相应的投资组合推荐,以提供给普通投资者更为切合自身实际情况的推荐指导。然而,股票的价格是一种动态的高噪声序列,影响其变化的因素是一个较为复杂以及随机的组合,因此对其进行有效预测是一项较为复杂、富有挑战的任务。与此同时,不同投资者千差万别的自身特性也给定制化的股票推荐带来了新的挑战。

目前学术界以及金融界主流的做法,主要是基于上市公司自身运营状况以及经济参考因素,如美元指数,通货膨胀指数(CPI)等,进行股票走势的预测。然而这种做法却忽视了个体投资者自身的特性,会导致投资者无法得到最符合其投资习惯的投资组合推荐。例如,稳健性投资者往往是风险厌恶型投资者,因此不适宜对其推荐高风险高回报的股票投资组合。同时也有主流推荐算法根据个体用户的个人信息以及过往投资喜好进行个性化的定制股票推荐,然而这种针对个体的推荐做法又过分依赖于个体自身的投资喜好以及过往记录,往往也不能给个体投资者最优的投资组合推荐。

众所周知,在金融市场中,机构投资者指的是从事证券投资的法人机构,主要有保险公司、养老基金和投资基金、证券公司、银行等。而机构投资者因其具有资深的投资经验以及金融背景,在金融市场中扮演着举足轻重的角色。因此本文针对上述挑战,基于高阶奇异分解,通过对股票自身特性以及广大机构投资者的投资行为进行挖掘学习,为广大普通投资者提供更加专业以及更贴合自身特性的投资组合推荐。

收稿日期:2014-03-20。茅斯佳,硕士生,主研领域:推荐系统。臧斌宇,教授。张谧,讲师。

本文最终选取 50 个机构投资者,即基金经理,在过往 20 个季度的投资数据进行分析测试,预测出机构投资者在两个季度的投资预期,为普通投资者进行了个性化的股票投资推荐。实验显示,个性化模型的准确率和收益率优于非个性化模型。

1　背景知识

1.1　高阶奇异值分解

高阶奇异值分解 HOSVD 是 Lieven 在文献 [1] 中首次提出的,文中 Lieven 介绍了 HOSVD 的算法,其应用的独特性以及与特征值分解的关系。在之后的诸多应用场景中,都将 HOSVD 方法作为基础进行数据挖掘。文献 [5] 是关于 HOSVD 算法的具体应用场景。文献 [2] 提出一种迭代张量高阶奇异值分解(HOSVD)图像缺失数据恢复方法。该方法首先利用拉格朗日乘子方法将张量核范数目标函数进行子问题分解操作,简化了求解过程,然后迭代地采用张量高阶奇异值分解阈值方法进行子问题求解,最终得到恢复后的图像缺失数据。而文献 [6,12,14] 是在高阶奇异值分解的基础上,提出了一种基于特征空间的快速张量分解算法。首先使用传统的子空间学习方法对观测图像进行降维,然后在低维的特征空间对训练数据进行张量分解。而文献 [13,15] 的研究着重于运用 HOSVD 的方法来解决标签推荐的问题。

上述的研究基本都是将高阶奇异值分解的方法用于特定的应用场景,这些应用大多将多阶张量一次性分解,而没有考虑其他更多的因素。本文加入了个性化特征信息。不同于之前其他研究,本文目的不是仅仅从静态角度分解矩阵或张量,而是着眼于为未来的投资决策做推荐。另外,HOSVD 将在不同阶段运用多次。本文提出的策略将从历史信息中提取特征值或特征信息,并用过去的历史信息预测股票未来的趋势。

1.2　股票预测

文献 [9] 认为基于账户的盈利操纵检测模型对于股票横截面收益有很强的非样本预测能力。文献 [10] 研究表明,个体投资者会投资那些信息公开透明的公司股票。此文同时提出,相比于信息闭塞的公司,信息公开透明的公司往往会为投资者带来更大的经济收益。

此外,很多研究关注于如何通过基本面分析,即通过财务数据指标分析来预测股票的收益。文献 [3,7] 选取上市满三年的证券公司作为研究样本,从财务指标与投资回报表现两个维度对上市证券公司的基本面表现进行科学的分析评价。首次验证在上市证券公司中从财务指标维度分析投资组合的收益可以跑赢大盘,这说明基本面分析,即本文中要提到的非个性化方法在证券股市场是有效的。文献 [4,11] 从证券市场参与者的有限理性出发,分析了基本面分析师在进行价格预测的过程中所面临的复杂制度环境,通过 MS(Microscopic Simulation)的建模方法与决策规则之间的相互作用建立了仿真的证券市场。文献 [8] 根据信息化背景下银行业的基本面信息和技术面信息,对银行股价的走势进行了分析并做出了预测。分析表明,银行业整体状况较好,银行股的股价也成上升趋势。

之前的研究着重于预测某只特定股票的未来涨跌情况,本文旨在通过探究基金经理未来可能做出的决策,为普通投资者提供更为全面的投资信息。另外,不同于其他方法只关注公开的股票信息,本文会从历史交易记录中提取个性化的特征信息以及结合投资者的个人特征进行推荐。

2　数据分析

2.1　数据简介

本文从雅虎财经(http://finance.yahoo.com)上获取数据,包括股票本身信息和诸多基金经理真实的交易数据。分析时将交易数据表示为一个三维立方体 $\{u,s,t\}$,u 表示用户维度即基金经理,s 表示股票维度,t 表示时间维度。数据集的股票总数为 2491 个,几乎包括了中国在上海和深圳两个证券交易所上市的所有股票,并且基金经理 u 的总数超过了 1000 人。为了便于演示,本文从不同类型的基金公司中选取了 50 个基金经理,基金公司类型涵盖了成长型基金、价值型基金等。至于时间维度,鉴于无法获得以日为单位的交易数据,这里将时间单位设定为季度。整个时间维度包含了从 2008 年 1 月份到 2012 年 12 月份的 20 个季度。

此外,收集的数据中也包含了每季度股票的价格,这些价格数据会在稍后被用于基本面分析来分析股票的表现,股票的表现通常用收益率来表示。

下面将使用大写字母序列 $(A,B,\cdots)$ 和 $(A',B',\cdots)$ 来分别表示张量和矩阵。一个有真实数据的 N 阶张量 A 表示,向量用粗体小写字母序列 $(a,b,\cdots)$ 表示,变量用小写字母序列 $(a,b,\cdots)$ 表示,集合用斜体大写字母序列 $(A,B,\cdots)$ 表示。

2.2　数据分析

分析时用一个向量来表示一个季度内基金经理的交易情况,这个向量的长度与股票的总数量相同。向量的每个元素表示该基金经理购买的相应股票的数量。购买的数量从数百股到数万股不等,平均数量约为 1000 股。

为了研究不同基金经理在某季度内购买股票行为的关联程度,这里向量表示各个基金经理在该季度购买股票的情况,通过计算不同向量之间的相关系数得到基金经理之间购买行为的相关度。基金经理的决策越趋同,相关系数越大,反正越小。相关系数的计算方式为: $c=(Cov(x,y))/(\sigma(x)\sigma(y))$,其中 x 和 y 为对应两位基金经理购买股票数量的向量。$Cov(x,y)$ 是两向量间的协方差,$\sigma(x)$ 和 $\sigma(y)$ 分别是两向量的方差。在数据集中,基金经理之间的平均相关系数是 -0.09,这意味着基金经理的购买行为略成负相关且接近于 0,表明了他们之间的投资行为较为独立。基金经理可能因为受到信息不对称、投资习惯、资金体量等的影响,从而导致完全不同的购买行为。因此,在向个人投资者提供建议时充分考虑每个基金经理的个性选择是有价值的。

为了进一步研究基金经理在不同季度的购买行为,也用向量表示某个特定基金经理,在不同季度下购买各只股票的数量,并计算这些向量之间的相关系数,得出平均相关系数是 0.61。所以,尽管基金经理在不同季度的购买行为是不同的,但是总体而言,每两个季度之间的购买行为呈现正相关性。这种规律标明经理在一段时期内投资决策不会有很大改变,这种规律为本文稍后提出的基于 HOSVD 的策略提供了基础,例如,可以从过去的交易数据中学习一种固有的投资模式,并且在这个模式的基础上预测未来的趋势。

然而,经常会出现不可测因素导致股票的突然变化,这将会造成两个连续季度之间呈现较小相关性甚至负相关性。这也就是本文后面将基于历史交易数据的个性化模型与基于股票本身特征的非个性化模型相结合的原因。

3 非个性化推荐—基于线性回归方法的非个性化推荐

本文针对普通投资者其个人信息，诸如个人喜好，历史投资记录等，同股票信息相结合，建立相应的多元线性回归模型，其中自变量为公司财报中的关键数据以及金融指标，因变量为股票评分，以学习得出个人对某只股票的喜好评分。其数学模型如下所示:

$$R_u = \sum_1^i \omega_i \times F_i \qquad \sum_1^i \omega_i = 1 \tag{1}$$

式中，F 表示影响个人购买该股票的因素，包括该股票相关信息，诸如股票所属行业等，以及个人历史交易记录等。R_u 表示最终个人对该股票的喜好评分。

非个性化策略的主要思想是通过这些经济指标预测股票在未来特定时期内涨跌的可能性。这些经济指标通常可以由基本面分析获得。基本面分析中最大的一部分是深入分析财务报表，也就是我们熟知的量化分析，它包括分析公司的收入、支出、资产、负债以及其他所有的金融指标。基本面分析师通过分析这些信息预测一个公司的未来表现。另外，财务报表分析可以将一系列财报项目结合起来分析总结，从而提前一年预测公司的收益变化。

根据文献[3]提到的经典理论，财务报表上的指标可以大致估计股票价格在未来的波动。我们采用回归分析这个统计过程来估计因变量之间的关系。更具体地说，回归分析可以帮助分析者了解在一些自变量固定不变，另一些自变量发生变化的情况下因变量变化的典型值。在回归分析中，估计目标是一个自变量函数时，称这个目标为回归函数。

一般的多元线性回归分析可以用来预测当任一指标变化时未来股价如何变化。一个正的变量意味着股价上涨，而一个负的变量意味着股价下跌。这个预测模型构成如下:

$$Pr(s) = \sum_{i=1}^{l} \omega_i \times f_i \quad \text{a. t.} \quad 0 \leqslant \omega_i \leqslant 1 \tag{2}$$

式中，f_i 是影响股价的指标，l 是指标的数量，$Pr(s)$ 是股价上涨的概率。

经营指标通常来源于历史财务报表或者当下的经营数据，包括全球经济状况、公司状况和交易状况等。本文从公司每季度披露的财务报表中选取三个公开可用指标，即每股收益、每股净资产和每股现金流作为本文模型的自变量。

每股收益(EPS)，又称每股税后利润、每股盈余，指税后利润与股本总数的比率。每股净资产是指股东权益与总股数的比率。其计算公式为: 每股净资产 =(总资产 - 负债) ÷总股数。每股现金流量是公司营业业务所带来的净现金流量减去优先股股利后与发行在外的普通股股数的比率。每股现金流量的计算公式如下: 每股现金流量 =(营业业务所带来的净现金流量 - 优先股股利) /流通在外的普通股股数。

对于每个股票，待预测的目标季度的实际每股收益率将作为训练的基准值。显而易见，高收益对应着高盈利可能性。

给定了基准值和指标后，ω_i 可以在线性回归中被学习。

当为用户做推荐时，市场中所有的基金经理根据 $Pr(s)$ 被分类，并且最好的 N 个基金经理被推荐给用户。由于 $Pr(s)$ 的预测仅基于发布的公开信息，所以为用户做的推荐是非个性化的。

在本文剩余部分，上述策略被称为 REG。

4 个性化推荐

本文认为向个人投资者提供个性化的建议是有价值的。投资习惯不仅由股票自身和经营指标决定，在很大程度上，投资习惯由从用户历史交易数据中挖掘出的特征因素决定，这是本文旨在探索的主旨。

非个性化的方法不考虑用户的偏好。本文提出先将目标用户设定为基金经理，运用个性化推荐方法，预测出基金预测的投资趋势。而后参考基金经理的投资趋势来为普通投资者提供建议。由于股票的表现很大程度上由大基金公司的操作决定，由此推断普通投资者跟随大户(基金经理)投资会获得更多收益。

在本文的剩余部分，这种策略被称为 HOSVD。

4.1 奇异值分解

奇异值分解 SVD(Singular Value Decomposition)是线性代数中一种将矩阵对角化的数值算法，并且被广泛的应用在信号处理、统计学等领域。其数学表示如下:

假设 F 是 I_1 I_2 矩阵，U 是 I_1 I_1 矩阵，其中 U 的列为 FF^{T} 的正交特征向量，V 为 I_2 I_2 矩阵，其中 V 的列为 $F^{\mathrm{T}}F$ 的正交特征向量，若 r 为 F 矩阵的秩，则存在奇异值分解:

$$F_{I_1 \times I_2} = U_{I_1 \times I_2} \cdot S_{I_1 \times I_2} \cdot V^{\mathrm{T}}_{I_2 \times I_2} \tag{3}$$

其中 FF^{T} 和 $F^{\mathrm{T}}F$ 的特征值相同，为 $\lambda_1 \cdots \lambda_r$，$S$ 为 I_1 I_2 矩阵，其中 $S_{ii} = \sqrt{\lambda_i}$，其余位置数值为 0，$S_{ii}$ 的值按大小降序排列。

用 F 乘以其转置矩阵 F^{T} 得:

$$FF^{\mathrm{T}} = USV^{\mathrm{T}}VSU^{\mathrm{T}} = US^2U^{\mathrm{T}} \tag{4}$$

奇异值分解的图形表示如图 1 所示。

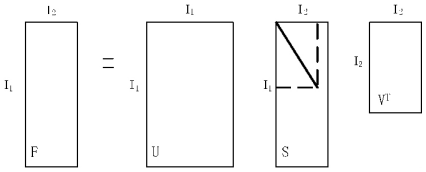

图 1 对矩阵 F 奇异分解的图形化表示

4.2 高阶奇异值分解算法

在现实生活中，遇到的计算场景往往需要处理超过两维的数据，这促使对原有基础的 SVD 分解算法加以改进。Lathauwer 等人[4]提出了 HOSVD 算法以解决高阶张量的分解问题。

首先，定义 N 阶张量 $A \in R^{I_1 \times \cdots \times I_N}$，以三阶张量为例，将张量 A 可以展开为如下三个矩阵:

$$A_1 \in R^{I_1 \times I_2 I_3} \quad A_2 \in R^{I_2 \times I_1 I_3} \quad A_3 \in R^{I_3 \times I_1 I_2} \tag{5}$$

式中，A_1、A_2、A_3 分别为对张量 A 中 I_1、I_2、I_3 三个维度进行展开操作后所的矩阵。其图形化表示如图 2 所示。

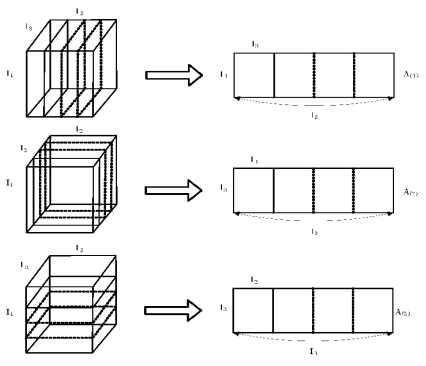

图 2 对三阶张量 A 高阶展开操作的图形化表示

对 A_1、A_2、A_3 分别进行 SVD 分解,得到:

$$A_1 = U^{(1)} S_1 V_1^{\mathrm{T}} \quad A_2 = U^{(2)} S_2 V_2^{\mathrm{T}} \quad A_3 = U^{(3)} S_3 V_3^{\mathrm{T}} \tag{6}$$

如图 3 所示。接下来对 S_1、S_2 和 S_3 降维。假设三个不同大小的 c_1、c_2 和 c_3 分别代表 S_1、S_2 和 S_3 降维后的维度,则降维后 $U^{(1)}$、$U^{(2)}$ 和 $U^{(3)}$ 的左奇异向量数量分别为 c_1、c_2 和 c_3。最后,运用降维后的 $U^{(1)}$、$U^{(2)}$ 和 $U^{(3)}$ 算得核心"特征张量"S。S 表示为:

$$S = A \times U_{c1}^{(1)\mathrm{T}} \times U_{c2}^{(2)\mathrm{T}} \times U_{c3}^{(3)\mathrm{T}} \tag{7}$$

最后,得到一个降维后的 A' 为:

$$A' = S \times U_{c1}^{(1)} \times U_{c2}^{(2)} \times U_{c3}^{(3)} \tag{8}$$

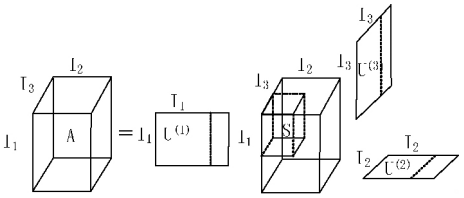

图 3　三阶张量 A 高阶奇异分解的图形化表示

4.3　基于高阶奇异值分解算法的推荐策略

该算法的目的是预测基金经理在接下来的一个季度会做何种交易(即基金经理预期购买的股票数量),以此为普通投资者提供投资信息。本文提出一种新的基于 HOSVD 的个性化推荐方法。假设一个交易张量 H,在这个策略中,每次交易 v 对应于一个三阶元组$\{b,s,t\}$,这个元组描绘了在时间 t 由基金经理 b 购买的股票 s 数量 v。本文的基本思想是从一段连续时间片中学习提取出"特征张量"。在一个特定时期内,假设这个特征张量在连续时间片内保持恒定,在假定的数据集下,这个假设大部分是成立的。具体地说,整个交易张量 H 是根据季度分片的,每个分片包含了所有基金经理本季度的交易数据。这样,在时间维度上,三阶张量就被分片成为一组二阶张量的矩阵。

为了对一个未来季度做出预测,首先,这个待预测季度的前三个连续季度需要被训练,即通过 HOSVD 多次分解重组获得"特征张量"。然后,基于这个特征张量对未来季度做出预测。这两步的详细步骤如下给出。

为了便于演示,下文设定待预测季度为 Q_4(任何一个待预测季度之前需要有至少三个季度)并且设定待预测季度之前的三个连续季度分别为 Q_1、Q_2 和 Q_3。

4.4　特征三阶张量构造

本文处理的三阶张量的三维分别是基金经理、股票数量和时间。分解因式阶段的目的不在于降阶,而在于学习核心特征张量并在下一步骤中使用它。

假设张量 A 由时间片 Q_1、Q_2 和 Q_3 组成,也可以认为是来源于整个张量 H 的一个片段,然后将 HOSVD 应用于张量 A。

首先进行展开,得到:

$$A_1 \in R^{User\times StockTime} \quad A_2 \in R^{User\times StockTime} \quad A_3 \in R^{User\times StockTime} \tag{9}$$

然后对 A_1、A_2、A_3 三个展开矩阵分别运用 SVD,每一个展开矩阵都形成三个新的矩阵。分别表示为:

$$A_1 = U_A^{(1)} S_{A1} V_{A1}^{\mathrm{T}} \quad A_2 = U_A^{(2)} S_{A2} V_{A2}^{\mathrm{T}} \quad A_3 = U_A^{(3)} S_{A3} V_{A3}^{\mathrm{T}} \tag{10}$$

然后用式(8)中得到的结果确定核心张量 S_A 为:

$$S_A = A \times U_A^{(1)\mathrm{T}} \times U_A^{(2)\mathrm{T}} \times U_A^{(3)\mathrm{T}} \tag{11}$$

本文提出的股票推荐模型认为核心张量 S_A 反映了基金经理、股票数量和时间之间的联系,因此将 S_A 用作特征张量。

上述过程如图 4 所示。通过对张量 A 应用 HOSVD,可得一个特征张量 S_A 以及三个矩阵 $U_A^{(1)}$,$U_A^{(2)}$ 和 $U_A^{(3)}$。如图 4 所示,这三个张量分别与用户(基金经理)、股票数量和时间有关。

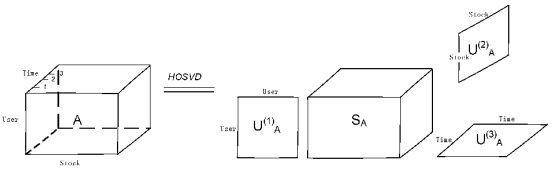

图 4　对张量 A 进行 HOSVD

4.5　下季度预测分析

再考虑另一个张量 B,这个张量由包含待预测季度的时间片 Q_2、Q_3 和 Q_4 组成。注意,张量 B 和张量 A 只有一个时间片的差别 $-Q_1$ 和 Q_4。他们在不同的时间序列中共享时间片 Q_2 和 Q_3,这说明张量 A 与 B 在某种程度上有延续性。如图 5 所示。

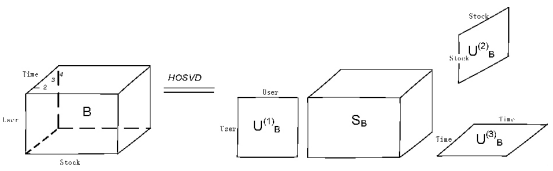

图 5　对张量 B 进行 HOSVD

由于 Q_4 为待预测时间片,因此此时间片中各个基金经理对所有股票的购买量是未知的。为了对张量 B 进行因式分解,将时间片 Q_4 中的股票数量设定为平均股票数量,这个平均股票数量作为所有股票数量的初始估值。然后按照图 5 所示的方式对张量 B 进行 HOSVD。展开操作如下:

$$B_1 \in R^{User\times StockTime} \quad B_2 \in R^{User\times StockTime} \quad B_3 \in R^{User\times StockTime} \tag{12}$$

得到展开矩阵 B_1、B_2,和 B_3 后,分别进行如下的 SVD:

$$B_1 = U_B^{(1)} S_{B1} V_{B1}^{\mathrm{T}} \quad B_2 = U_B^{(2)} S_{B2} V_{B2}^{\mathrm{T}} \quad B_3 = U_B^{(3)} S_{B3} V_{B3}^{\mathrm{T}} \tag{13}$$

下面将 HOSVD 逆向操作以获得估计张量 B',方法与式(8)类似,然而,这里并不是使用压缩核心张量和矩阵的方式来构建估计张量,而是使用从张量 A 中获得的特征张量以及从张量 B 中获得的各个维度矩阵,构建形式如下:

$$B' = S_A \times U_B^{(1)} \times U_B^{(2)} \times U_B^{(3)} \tag{14}$$

图 6 描绘了这个构建过程。通过这种方法 Q_4 中的初始平均数量被替换成了更为精确的预估值。这个预测同样可以认为是基金经理购买这些股票的可能性。

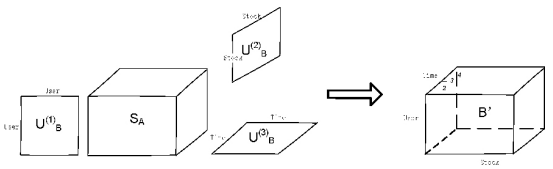

图 6　对张量 B 的反向操作

反向操作的想法是以待预测季度与之前连续季度之间没有突然变化为基础的,在这种情况下,特征张量可以表示连续时间片一致特征。然而,市场中总会存在一些不可测因素,此时连续时间片不再表现出明显的继承性,此时用特征向量抽象提取一段时间内购买特征的方法失效,这也是本文提出将上述方法与本文第 3 节中所述的非个性化回归方法相结合的原因。

4.6　基于预测分析的推荐

在获得对下一季度的预测后,需要为用户做出推荐,这里应用了两种方法做推荐。

4.6.1　AVE 方法

第一种方法叫做平均法(AVE)。在这种方法中,针对每一只股票,将所有基金经理购买的数量取平均数。然后将这些股票按照购买平均数进行分类,取购买数量最多的 N 支股票向用户做推荐。

4.6.2 BST 方法

第二种方法称为最优法(BST)。不同于 AVE 法,BST 根据一个在过去的三个季度内获利最大的基金经理的购买数量进行推荐。

具体来说,像之前提到的一样,数据集中包含了过去季度内的每只股票价格,以及基金经理在过去季度内的购买行为(也就是说他们买的股票名称及数量)。然后,基于上述两个数据源,基金经理的表现可以通过计算平均收益率获得。

首先,股票 i 在季度 j 内的收益率 e_j^i 可以表示为:

$$e_i^j = \frac{p_i^j - p_i^{j-1}}{p_i^{j-1}} \tag{15}$$

式中,p_i^j 是股票 i 在季度 j 内的价格,显然,p_j^{i-1} 就是上一季度的价格。

然后,基金经理 k 在季度 j 内的表现可以按如下公式计算:

$$e_k^j = \frac{\sum_{i \in I_k^j} v_i \times e_i^j}{\sum_{i \in I_k^j} v_i} \tag{16}$$

式中,I_k^j 是基金经理 k 在季度 j 内购买的一系列股票。e_k^j 是基金经理的平均收益率。

显然,高的值意味着该基金经理在过去做了非常成功的投资,而该基金经理的专业意见也可以成为对其他投资者的专家建议。在本文的实验中 e_k^j 是待预测季度的前三个季度内的平均收益率,这样就将基金经理在过去三个季度内的购买行为综合考虑了,而非单一参照一个季度。

当挑选出获得最高平均收益率的基金经理之后,预测该基金经理在接下来一个季度的购买行为,并推荐给用户。具体来说,将预测的购买股票数量按降序排列,并选取最多的 N 只股票推荐给投资者。与此同时,给出该基金经理被推荐的股票数量占其购买总数的比例。

4.7 与线性回归模型的整合

就如 2.2 节分析的一样,当股票市场变化较为平稳时,意味着前后两个季度具有正相关性,此时个性化推荐方法较为有效;但当股票市场波动较为剧烈的时候,由于基本面分析只关注于财务指标的变化,此时非个性化推荐方法较为有效。为了更为有效地解决现实场景中复杂的问题,这里将这两种方法相结合,即赋予非个性化推荐的线性回归方法和个性化推荐的高阶奇异值分解算法权重,分别为 w 和 $1-w$。

这种方法的基本思想是将两种方法的推荐结果予以加权。这两种方法均可以得到一个推荐股票的排序,按照其推荐程度由高到低进行排列。

本部分提出的整合方法赋予两种方法各一个权重,这个权重决定了两种方法各自对最终结果的贡献大小,这个整合模型的公式如下:

$$r_i - wr_i^{HOSVD} + (1-w) r_k^{REG} \quad \text{s.t.} \quad 0 \leqslant w \leqslant 1 \tag{17}$$

式中,r_i^{HOSVD} 和 r_k^{REG} 表示 HOSVD 策略和 REG 策略所得的排序序列,r_i 直观地由两种策略加权整合得到。

这种整合的策略既考虑了财务指标这类非个性化因素,又考虑了用户的历史交易所反映出的个性化特征。通过调整 w,可以让整合的策略适应不同的应用环境。当市场波动较为剧烈时,可加大非个性化策略的权重;当市场较为平稳时,可加大个性化策略权重。市场的波动程度可用当前时间片与之前时间片的相关系数来界定。权重不同时,整合策略的结果也会有较大差异。

最后,该方法会推荐整合后综合排名前 N 个股票。W 越大,个性化推荐的程度越高,在本文的后面,用 COM 来表示这种整合的方法。

5 实验评估

本文最终从雅虎财经数据库中选取 50 个基金经理在过去 20 个季度所披露的基金数据作为基础数据进行迭代学习。具体实验数据分析见第二部分。

5.1 评估方法

本文的实验数据分为两类,第一类数据是基金经理购买的股票数量,这类数据将作为基准值来评判本文推荐算法的表现情况;第二类数据是根据本文推荐算法购买未来股票的收益率。第二类基准数据是目标季度每只股票的实际收益,将所有股票按照实际收益率的大小降序排列,并由这个降序序列产生测试集 T。

在两种情况下定义命中事件,第一种情况,测试集 T 由当季中所有收益为正的股票组成,此时,命中事件表示推荐的某只股票确实是可盈利的;第二种情况,测试集 T 也可由降序序列中前 k 只股票组成,即收益率最大的前 k 只股票,此时命中事件表示推荐的股票在测试集 T 中。推荐的准确度可由准确率与召回率来测定。

5.1.1 基于平均绝对误差的评估

首先,预测的准确性可以由 MAE 来评价。MAE 定义如下:

$$MAE = \sum_{u \in U} \sum_{i \in I} (\hat{r}_{ui} - t_{ui}) \tag{18}$$

其中 U 是基金经理的集合,I 是股票的集合,$\hat{r}_{ui}$ 是预测值,r_{ui} 是真实数量,MAE 给出了平均绝对误差。MAE 越大,表示误差越大。在 HOSVD 方法中,最终推荐集合的形成的基础是预测集,所以,预测准确度是本文实验中应关注的焦点之一。

5.1.2 基于准确率与召回率的评估

其次,推荐的准确率还可以由准确率与召回率来评估:

$$Precision = \frac{|R \cap T|}{|R|} Recall = \frac{|R \cap T|}{|T|} \tag{19}$$

式中,$Precision$ 表示准确率,$Recall$ 表示召回率。R 表示推荐的股票集合,而 T 是由收益率大于 0 或收益率最高的那部分股票组成。

5.1.3 基于收益率的评估

收益率描述的是如果用户按照不同的投资策略投资可以获得的实际收益,因此,这是最重要的评价推荐质量的标准。

可用式(15)计算推荐集合 R 中每只股票的收益率,这个收益率计算的是某只股票当前季度的价格较前一季度价格的变化率。而式(16)可用来计算一个推荐集合中股票的平均收益率。具体地,在 REG 模型中,假设推荐集合中每只股票的购买量相同;而在 HOSVD 模型中,假设每只股票的购买数量与预测值成正比,即预测可能性越大,购买量越大。

5.1.4 基于推荐多样性的评估

多样性描述的是推荐结果的新鲜度。在这里我们选取的基准推荐是第三节中介绍的非个性化推荐。本文定义新鲜度评估标准如下:

$$Novel = \frac{BP \cap P}{|NP|} \tag{20}$$

式中，*Novel* 表示推荐的多样性程度，*NP* 表示非个性化推荐策略的推荐结果，*P* 表示个性化推荐的推荐结果，$NP \cap P$ 表示两种推荐策略下，推荐结果的交集，$|NP \cap P|$ 表示此交集中元素的数量。$|NP|$ 表示非个性化推荐策略的推荐集合中的元素个数。

当 *Novel* 值越低时，推荐结果的多样性程度越高; 反之越低。

5.2　实验结果

下面给出三种推荐策略的实验结果，分别是线性回归方法，高阶奇异值分解法和两种方法的结合策略，并且用准确率、召回率和收益率三个指标对它们进行评价。另外，对于高阶奇异值分解法，专门用 MAE 指标来评价其准确率。

本文选取两个独立的季度作为测试集来检测学习结果，用 Q_a 和 Q_b 来分别表示这两个时间片。其中，Q_a 与前一个季度的相关系数是 0.65，这说明 Q_a 与前一季度的相关性较高，此时市场没有发生突变，较为平稳; 而 Q_b 与前一个季度的相关系数是 −0.06，说明 Q_b 与前一季度关联性较小，此时市场发生突变，波动较大。

5.2.1　预测准确率表现

本文用 MAE 方法来评估 HOSVD 方法的准确率，当使用 MAE 方法时，需要首先对购买数量进行归一化。实验结果表示，用 HOSVDS 方法的所有推荐结果的平均 MAE 为 0.0612，这说明 HOSVD 方法的推荐准确定度较高。

5.2.2　推荐准确率表现

推荐结果的准确率可用准确率和召回率表示。这里把推荐集的大小分别设定为 5、10、15 和 20。每一种方法都会推荐表现最好的 *N* 只股票给用户。值得注意的是，由于测试集 *T* 是由所有收益率大于 0 的股票组成的集合，因此 *T* 集合的元素数量很多，召回率的数量级非常小。

在运用 HOSVD 方法得到 Q_a 和 Q_b 这两个时间片后，分别运用两种基于预测的方法进行最终推荐。这两种方法是平均法(AVE) 和最优法(BST)。实验结果见表 1 所示，将高阶奇异值分解算法(HOSVD)、线性回归算法(REG) 及两种方法结合的算法(COM) 进行了比较。有趣的是，对于 Q_a 和 Q_b，这三种方法的相对优劣并不一致。对于 Q_a，HOSVD 算法明显优于 REG，而对于 Q_b，REG 算法的结果比 HOSVD 更优。这个结果说明在市场较为稳定，即待测季度与之前相关系数较高时，HOSVD 方法的效果较好; 而在市场波动剧烈时候，基于财务指标，即基本面分析的线性回归方法的效果较好。COM 算法的结果介于上述两种算法之间。图 7 描述了当 *w* 设定为不同值时，Q_a 和 Q_b 的变化。很明显，对于 Q_a 而言，当 *w* 设定为 0.8 时，推荐结果较好; 对于 Q_b 而言，当 *w* 设定为 0.2 时，推荐结果较好。总体而言，运用 HOSVD 方法后，进行最终推荐时，使用 BST 方法的表现要优于使用 AVE 方法。这表明了投资者的策略应追随那些表现最好的基金经理，而不是依据所有基金经理的平均购买行为进行投资。由于测试集 *T* 的元素数量是固定的，因此召回率的变化与准确率的变化一致，在此不做特别说明。

表 1　策略的预测准确率

Qa 准确率	REG	HOSVD		COM			
		AVE	BST	w = 0.2	w = 0.4	w = 0.6	w = 0.8
Top 5	0.7	0.78	0.79	0.7	0.76	0.77	0.73
Top 10	0.6	0.82	0.85	0.63	0.7	0.76	0.7
Top 15	0.59	0.77	0.86	0.61	0.76	0.79	0.67
Top 20	0.51	0.79	0.86	0.58	0.72	0.77	0.64

续表 1

Qb 准确率	REG	HOSVD		COM			
		AVE	BST	w = 0.2	w = 0.4	w = 0.6	w = 0.8
Top 5	0.73	0.4	0.43	0.57	0.54	0.48	0.46
Top 10	0.65	0.39	0.42	0.56	0.53	0.45	0.43
Top 15	0.6	0.38	0.41	0.53	0.51	0.44	0.43
Top 20	0.56	0.37	0.38	0.48	0.48	0.4	0.41

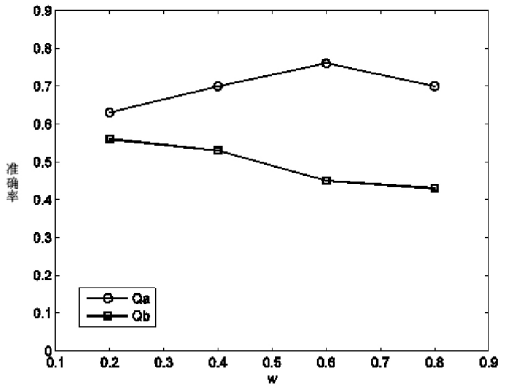

图 7　预测准确率与 *w* 的关系

如图 7 所示，当测试集为 Q_a 时，即市场较为平稳时，*w* 为 0.6 实验能取得最高准确率，这说明个性化推荐方法此时能取得较好结果，可加大个性化策略权重; 当测试集为 Q_b 时，即市场波动较为剧烈时，*w* 为 0.2 实验能取得最高准确率，这说明个性化推荐方法此时不能取得较好结果，而增加非个性化推荐方法能提高准确率，因此可加大非个性化策略的权重。

5.2.3　收益率表现

在得到推荐集合后，HOSVD 方法和 REG 方法的收益率可以通过上述方法获得，如表 2 所示。

表 2　收益率

Q_a 收益率	REG (%)	HOSVD		COM			
		AVE (%)	BST (%)	w = 0.2 (%)	w = 0.4 (%)	w = 0.6 (%)	w = 0.8 (%)
Top 5	5.59	7.36	7.53	6.35	7.77	7.75	6.18
Top 10	5.43	6.99	7.19	6.29	7.64	7.67	6.16
Top 15	5.39	6.77	6.94	6.21	7.61	7.62	6.15
Top 20	5.19	6.59	6.74	6.17	7.58	7.66	6.11
Q_b 收益率	REG (%)	HOSVD		COM			
		AVE (%)	BST (%)	w = 0.2 (%)	w = 0.4 (%)	w = 0.6 (%)	w = 0.8 (%)
Top 5	5.41	2.33	2.34	4.45	3.41	3.37	3.27
Top 10	5.35	2.19	2.24	4.37	3.39	3.31	3.21
Top 15	5.19	2.16	2.21	4.31	3.32	3.27	3.18
Top 20	5.03	2.14	2.19	4.28	3.26	3.24	3.15

从表 2 中可以看出，对 Q_a 而言，两种方法的结合(COM) 比任意单独一种方法效果要好，由此可见，尽管 HOSVD 算法准确率最高，但最终推荐的股票集合不一定是收益率最高的。对于 Q_b 而言，REG 方法给出的推荐结果最好，而 COM 方法给出的结果介于 REG 与 HOSVD 之间，由此可见，无论市场稳定与否，COM 方法作为一种稳健的算法，可以提供最“安全”的投资策

略。图 8 展示当 N 取 10 时，收益率与 w 变化的关系。与图 7 中的准确率比较后可以发现，一般而言，较高的准确率对应较高的收益率。

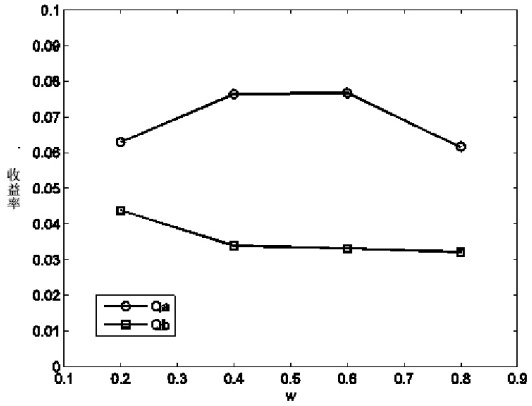

图 8 收益率与 w 的关系

如图 8 所示，当测试集为 Q_a 时，w 取 0.4 至 0.6 可以取得较高的收益率；当测试集为 Q_b 时，w 取 0.4 可以取得较高收益率。

5.2.4 多样性表现

分别用非个性化策略和个性化策略得到两个推荐列表以及每个推荐对应的推荐力度。当选取不同的推荐集合大小时，会产生不同的推荐集合，个性化策略的推荐多样性表现可通过式(20)算得，实验结果如表 3 所示。

表 3 多样性

多样性	Top 5(%)	Top 10(%)	Top 15(%)	Top 20(%)
Q_a	5.5	7.1	8.9	11.2
Q_b	4.1	6.2	8.2	10.4

从表 3 中可以看出，与传统的非个性化策略相比，本文提出的个性化推荐策略的新鲜度较高，即推荐的多样性较强。

6 结 语

本文认为，对股票的分析研究除了着眼于传统的基本面分析之外，对基金经理投资模式分析与预测可以为个人投资者提供更多关于股票未来趋势的信息。

本文将传统基本面分析即非个性化推荐与对基金经理投资模式的分析预测相结合，为投资者提供更多个性化信息。用多元线性回归分析的方法构建了非个性化模型，用高阶奇异值分解的方法构造个性化模型，并将上述两个模型加权整合，以获得在不同市场环境下的稳健推荐算法。

本文通过实验对比了个性化模型与非个性化模型的推荐效果，结果显示，个性化模型的准确率和收益率优于非个性化模型，且个性化模型推荐结果的多样性程度较高。

参 考 文 献

[1] De Lathauwer L, De Moor B, Vandewalle J. A multilinear singular value decomposition [J]. SIAM journal on Matrix Analysis and Applications, 2000, 21(4): 1253-1278.

[2] 周俊秀，裘国永，刘侍刚，等. 迭代张量高阶奇异值分解的图像恢复方法 [J]. 计算机应用研究，2013，30(11)：3488-3491.

[3] Wahlen J M, Wieland M M. Can financial statement analysis beat consensus analysts' recommendations [J]. Review of Accounting Studies, 2011, 16(1): 89-115.

[4] Zack, Gerard M. Financial Statement Analysis [J]. Financial Statement Fraud: Strategies for Detection and Investigation, 2013: 209-213.

[5] Wei C, Hsu W, Lee M L. A unified framework for recommendations based on quaternary semantic analysis [C] //Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval. ACM, 2011: 1023-1032.

[6] Rovid A, Szeidl L, Varlaki P. Data representation in HOSVD-DCT based domain [C] //Intelligent Engineering Systems (INES), 2013 IEEE 17th International Conference on. IEEE, 2013: 103-106.

[7] 齐岳，孙彬，崔喜君. 基于基本面的券商股投资组合跑赢大盘的可能性分析 [J]. 经济问题，2013(10)：45-50.

[8] Bowen R M, Burgstahler D, Daley L A. Evidence on the relationships between earnings and various measures of cash flow [J]. Accounting Review, 1986: 713-725.

[9] Bowen R M, Burgstahler D, Daley L A. The incremental information content of accrual versus cash flows [J]. Accounting Review, 1987: 723-747.

[10] 朴军. 基于 MS 建模方法的基本面分析师预测行为研究 [J]. 系统仿真学报，2006，18(3)：594-596.

[11] Li Q, Shi X, Schonfeld D. Robust HOSVD-based higher-order data indexing and retrieval [J]. IEEE SIGNAL PROCESSING LETTERS, 2013: 1-1.

[12] Wang H, Ahuja N. A tensor approximation approach to dimensionality reduction [J]. International Journal of Computer Vision, 2008, 76(3): 217-229.

[13] 王雅峰，丁彦蕊. 一种快速张量分解算法及其在人脸识别中的应用 [J]. 电子设计工程，2013，21(19)：30-32.

[14] Xu Y, Zhang L, Liu W. Cubic analysis of social bookmarking for personalized recommendation [M] //Frontiers of WWW Research and Development-APWeb 2006. Springer Berlin Heidelberg, 2006: 733-738.

[15] Luo D, Ding C H Q, Huang H. Multi-Level Cluster Indicator Decompositions of Matrices and Tensors [C] //AAAI. 2012.

(上接第 278 页)

[2] 李水友，刘智勇. 城市交通感应控制综述 [J]. 城市交通，2006，4(6)：64-69.

[3] 卢凯，徐建闽，郑淑鉴. 交通信号协调控制方案过渡优化算法 [J]. 交通运输工程学报，2012，12(6)：97-103.

[4] 解菲. 浅析我国城市交通管理与控制 [J]. 中国市场，2008，43(15)：18-19.

[5] 刘小明，王飞跃. 基于 Agent 的区域交通流协调控制的研究 [J]. 计算机工程，2003，29(9)：45-47.

[6] 董友球，刘智勇. 基于 Q 学习的区域交通控制方法 [J]. 五邑大学学报(自然科学版)，2008，22(2)：15-18.

[7] WieringM, Vreeken J, Veenen J V, et al. Simulation and Optimtzation of Fraffic in a City [C] //Proceedings of IEEE Intelligent Vehicle symposium (IV04), Italy, June 14-17, 2004.

[8] 李洪中. 基于模糊控制的智能交通灯系统的研究与设计 [D]. 兰州：兰州交通大学，2013.