文章编号:1672-2477(2020)04-0001-13

在金融科技中基于人工智能算法的风险特征因子筛选框架的建立和在期货价格趋势预测相关的特征因子刻画的应用

袁先智[1,2,3],周云鹏[3*],刘海洋[3*],严诚幸[3*],
钱国骐[4],钱晓松[5],汪冬华[6],李志勇[7],
李祥林[8],林健武[9],沈思丞[3],曾 途[3]

(1.成都大学 商学院,四川 成都 610106;2.上海立信会计金融学院 金融科技学院,上海 201620;
3.成都数联铭品科技有限公司,四川 成都 610000;4.墨尔本大学 数学与统计学院,澳大利亚 墨尔本 VIC3010;
5.苏州大学 金融工程研究中心,江苏 苏州 215000;6.华东理工大学 商学院,上海 200237;
7.西南财经大学 金融学院,四川 成都 611137;8.上海高级金融学院,上海 200030;
9.清华大学 深圳国际研究生院,广东 深圳 518057)

摘要:研究的目的是建立对影响大宗商品期货价格变化趋势的关联风险特征因子的提取框架和配套的推断逻辑原理。具体来讲,以金融科技中大数据概念为出发点,利用人工智能中的吉布斯随机搜索(Gibbs Sampling)算法为工具,全面地陈述如何提取高度关联大宗商品期货价格变化的风险特征因子的流程和配套的逻辑原理,即采用(在马尔科夫链蒙特卡洛(MCMC) 框架下)人工智能中的吉布斯随机抽样算法,结合 OR 值(Odds Ratio)作为关联分类和验证标准,实现从大量风险因子的数据中提取与大宗商品期货(铜)价格趋势变化相关的特征因子并进行分类,从而可用于构建支持期货价格趋势变化分析的特征指标。实证分析结果表明,该特征提取方法能够比较有效地刻画大宗商品期货(铜)价格的趋势变化,为业界进行大宗期货交易和风险对冲的管理提供了一种新的分析维度。另外,从影响价格趋势变化的特征因子中筛选出高度关联的特征指标的大数据分析方法,是与过去文献中对价格趋势分析的不同之处和创新点。

关 键 词:大数据; 吉布斯(Gibbs)随机搜索算法; 特征筛选; 关联方; 价格趋势变化

中图分类号:F803.9;O212.2;TP182 **文献标识码**:A

期货市场具有规避风险和价格发现的功能。期货价格是期货交易的核心要素之一,同时也是反映整个期货市场运行状况的主要因素,合理有效的期货价格可以起到先导作用并弥补现货价格的滞后。随着国内期货市场的不断完善和发展,期货市场在宏观经济运行中的作用也显得愈发突出,因此对于期货定价的研究具有重要意义。

大宗商品作为期货市场的主要标的,“消费属性”是大宗商品的基本属性。不过,随着金融市场的不断发展以及期货期权、商品 ETF 等金融产品的不断丰富,大宗商品的金融化特征不断加强[1-2],除传统意义下的“微观”和“宏观(基本面)”等因素以外,更多因素对大宗商品的价格产生影响开始愈发明显,其中,“微观因素”与“宏观因素”包括经济发展对商品的需求、生产技术、地缘政治,以及事件风险等因素。更多因素包括大宗商品市场与其他金融市场间的价格协整关系,资本市场和国际货币政策的变动,国际投机力量以及资本的流动性等。同时,由于大数据时代带来的信息的量级递增,许多高度关联大宗商品价格的关联因素以非结构化数据的形式出现,对大数据的分析处理正成为解决和分析传统问题盲区的有效手段。因此在大数据框架下,以海量的结构化和非结构化的大宗商品数据为基础数据池,通过大数据特征筛选和提取

收稿日期:2020-01-10
基金项目:国家自然科学基金资助项目(U1811462;71971031)
作者简介:袁先智(1965-),男,重庆人,教授,博士.
通讯作者:周云鹏(1993-),男,云南昆明人,金融科技工程师,硕士.
刘海洋(1995-),男,浙江文成人,金融工程师,硕士.
严诚幸(1985-),男,四川青川人,金融工程师/讲师,硕士.

方法，建立关于大宗商品价格（变化）趋势的影响因素研究，尝试建立在给定的误差容忍度下与大宗商品价格（趋势）变化高度关联的特征风险因子。这类通过大数据框架筛选提取出的高度关联的风险因子，将帮助和改善针对大宗商品价格趋势变化的解释能力。

到目前为止，尽管大多数的大宗商品定价模型能够在很大程度上拟合期货价格的期限结构以及价格变化规律，但是这些传统模型对于期货价格变化的解释还存在许多问题，比如不能很好地反映所有相关指标对期货价格的影响：一个基本的原因在于目前的大宗商品定价模型只是基于传统的结构化数据信息，通过因果关系来对商品价格的变化进行描述。但是，大宗商品的“消费、金融二重属性”给期货价格变化所带来的影响是复杂的，传统的定价模型无法反映海量的非结构化数据提供的相关信息。Yuan[3]等通过大数据全息画像方法（也叫“Hologram”方法）实现了结构化数据与非结构化数据的实时融合，然后完成针对中小微企业信用评分的全息画像评估方法与平台的建立，实现对中小微企业的风险特征的刻画和特征风险因子的提取，用于建立针对中小型企业贷款的信用评估能力。我们本文研究的目的是建立对影响大宗商品期货价格变化趋势的高度关联的风险特征因子的提取框架。具体来讲，以金融科技中大数据概念为出发点，利用人工智能中的吉布斯随机搜索（Gibbs Sampling）算法为工具，全面陈述了如何提取高度关联大宗商品期货价格变化的风险特征因子的流程和配套的逻辑原理，即采用（在马尔科夫链蒙特卡洛（MCMC）框架下）人工智能中的吉布斯随机抽样算法，结合 OR 值（Odds Ratio，又称为“比值比”或“优势比”，参见“附录 1”的描述）作为验证标准，实现从大量风险因子的数据中提取与大宗商品期货（铜）价格趋势变化相关的特征因子并进行分类，从而用于构建支持期货价格趋势变化分析的特征指标。实证分析结果表明，研究讨论的特征提取方法能够比较有效地刻画大宗商品期货（铜）价格的变化趋势，这为业界进行大宗期货交易和风险对冲的管理提供了一种新的分析维度。另外，此次研究讨论的从影响价格趋势变化的特征因子中筛选出高度关联的特征指标的大数据分析方法，这是与过去文献中对价格趋势分析的不同之处和创新点。

基于上述的介绍，研究的工作重点是在给定的误差容忍度标准下，提炼出与大宗商品价格变化具有高度相关的特征指标，围绕大宗商品的基础指标、产业指标、宏观指标等指标建立一套完善的大宗商品价格特征指标体系，为预测大宗商品的价格变化，提供一种基于大数据思维的全新方法。

研究以宏观及微观的因素为元素，以大数据框架下大宗商品期货相关的海量非结构化数据和结构化数据作为基础的超过 126 个与期货价格变化相关的因子作为初步风险因子数据池（参见“附录 2”中的陈述），利用吉布斯随机搜索（Gibbs Sampling）算法构建关于影响大宗商品期货价格变化的大数据特征筛选方法，完成对大宗商品期货价格变换的特征提取，然后以 Logistic 回归模型为工具构建基于特征集的特征权重比照，形成对期货定价影响强弱的价格风险因子特征的排序，最终进行实证检验。实证结果表明大数据特征提取算法能够有效地提取刻画沪铜指数合约价格趋势的特征，这些特征包含基础特征、消费市场特征和宏观经济指标多个维度，并支持我们实现对期货价格变化较好的预测性。

1 相关文献和相关研究综述

随着期货市场的不断发展与完善，对于期货价格的研究也成为学术界的重点研究领域。目前国内外对于大宗商品期货价格的研究主要集中在布朗运动模型及其扩展模型，大宗商品期货价格波动率，不同金融市场与期货价格之间联动性三个方面的研究。

在布朗运动模型及其扩展模型为主的期货价格研究方面，Brennan[4]等假设商品现货价格服从布朗运动，并提出以现货价格为状态变量的单因素模型。Schwartz[5]提出了以现货价格和随机便利收益为状态变量的二因素模型，同时又以利率作为第三个状态变量提出了三因素模型。Cassassus[6]等在 Schwartz 三因素模型的基础上提出了基于三因素仿射模型的仿射期限结构模型。王苏生[7]等在 Schwartz[8]等的基础上提出了以短期偏离、中期偏离和长期均衡为状态变量的三因素模型。韩立岩[9]等对能源商品期货也进行了研究，提出了以现货价格、便利收益和长期收益为状态变量的三因素期限结构模型，杨胜刚[10]和朱晋[11]也对基于三因素结构来研究期货价格与其他因素的关系提出了看法。

在针对大宗商品期货价格波动率变动的研究方面，张保银[12]、董珊珊[13]、黄健柏[14]等分别通过建立 VEC 模型、分数协整向量自回归模型（FC-VAR）、状态空间模型，基于实证分析认为我国沪铜商品期货价

格波动具有尖峰厚尾、集聚性和长记忆性等特征。Hamilton[15]等将马尔科夫链引入自回归模型中，提出了 RS-ARCH 模型。

在针对不同金融市场与期货价格之间联动性的研究方面，高辉[16]、张屹山[17]和郭树华[18]等分别用格兰杰检验、协整分析、误差修正模型（ECM）等计量方法分析了国内外金属期货市场之间的价格联动性。胡东滨[19]等运用 DCC-GARCH 模型对于金属期货与外汇、货币市场的动态相关性进行了深入研究。Yue[20]等采用 VAR-DCC-GARCH 模型，研究中国金属市场和 LME 市场金属价格间的动态联动性。李洁[21]等考虑不同期货市场、不同期货品种间的关联关系，对中英期货市场的期铜、期铝、期锌之间的价格交叉影响和风险进行了针对性地分析。

在过去几年的研究中，针对金融市场和 FOF 等产品方面，袁先智[22]等完成了针对 FOF 和中小微企业风险关联特征的提炼和刻画（特别是对于非结构化指标的提炼和刻画），并应用于金融业界实践。本文研究的目标是建立对影响大宗商品期货价格变化趋势的高度关联的风险特征因子的提取框架和配套的实践流程。研究结果显示，与传统的计量分析方法相比，大数据特征提取方法更能有效地在高维度的特征空间中对商品期货价格变化趋势特征进行特征刻画，同时利用大数据特征提取得到的特征集合建立针对沪铜价格变化的趋势分析，结果如表 1 所示。

2　基于人工智能算法的风险特征因子筛选框架的建立

研究以金融科技中大数据概念为出发点，利用人工智能中的吉布斯随机搜索（Gibbs Sampling）算法为工具，全面地陈述了如何提取高度关联大宗商品期货价格变化的风险特征因子的流程和配套的逻辑原理，即采用（在马尔科夫链蒙特卡洛（MCMC）框架下）人工智能中的吉布斯随机抽样算法，结合 OR 值（Odds Ratio，又称“比值比”或“优势比”）作为验证标准（参见“附录 1”的介绍），实现从大量风险因子的数据中提取与大宗商品期货（铜）价格趋势变化相关的特征因子并进行分类，从而用于构建支持期货价格趋势变化分析的特征指标。

2.1　基于吉布斯（Gibbs）随机搜索和 Logistic 回归模型方法的关联风险因子筛选算法思想

在人工智能算法中的吉布斯抽样（Gibbs Sampling）是一种简单有效并且广泛应用的马尔科夫链蒙特卡洛（MCMC）算法，特别适宜于从复杂的多元概率分布产生随机向量。考虑到某大宗商品价格涨跌受众多因素决定，而我们只有和其有潜在关联的 M 个因素（记为 $z_i, i=1,2,\cdots,M$）的观测值。因此假设价格涨跌是服从依赖于这 M 个（关联）因素的一个概率模型，那些影响价格涨跌但未被观测到的因素的影响全部被归入该概率模型的随机部分。目标是从这 M 个因素中发现和价格涨跌最关联的因素，要完成这一目标是一件非常具有挑战性的工作。由于 M 一般比较大，同时 M 个元素之间也互相关联（在大数据的概念下），不能孤立地进行处理。但考虑任何一组给定的因素和价格涨跌的可能的关联可以用 Logistic 回归模型为工具来进行分类处理和分析，基于这个基本的思路，可以构造一个与价格涨跌和任何一组因素关联关系的概率分布函数，这个分布函数可以表示为 $p(z):=p(z_1,\cdots,z_M)$，这里 z_i 是一个指标变量，$i=1,2,\cdots,M$；$z_i=1$ 表示关联因素 i 被用在所分析的 logistic 回归模型里，$z_i=0$ 表示因素 i 没有被用在所分析的 logistic 回归模型里，然后就可以把原始目标的寻找转换为找出基于 $(z_1,\cdots,z_M)$ 的最优值，使得 $p(z)$ 获得最大概率。因为搜索空间的大小是 2^M，一个指数增长的计算量级。因此直接找最优的 $(z_1,\cdots,z_M)$ 在计算上不可行（即所谓的“NP 问题”，参见文献[23]中的讨论）。通过把寻找最优值的问题转换为对应的概率分布 $p(z)$ 的最大值的估计工作，可以由吉布斯（Gibbs）随机抽样的搜索方法产生一系列随机向量来解决问题（因为最优 z 值通过随机样本的产生来实现），这样就可以在生成的随机样本里搜索最优关联因素集，解决现实的计算复杂度这种 NP 问题（参见文献[23]，文献[24]和相关文献的讨论）。

为了将所有与价格关联的（风险）因子以 Logistic 回归模型为工具进行关联关系的强弱分类，首先按照吉布斯抽样方法，把每个步骤涉及到的一个关联变量的值替换为以剩余变量的值为条件的条件概率，从这个条件概率分布中抽取对应变量的值，即将 z_i 替换为从条件概率分布 $p(z_i \mid z_{\setminus i})$ 中抽取的值（其中 z_i 表示 z 第 i 个元素的指标变量，$z_{\setminus i}$ 表示 $z_1,z_2,\cdots,z_M$ 去掉 z_i 这一项）。再按照一个特定的顺序在变量之间进行循环计算，每一步按照某个变量关于其他所有变量组合的条件概率分布随机地对该变量进行更新。

结合一个例子，在此对算法步骤实现进行一个简单的表述：对于一个三维向量变量 $(z_1^{(\tau)},z_2^{(\tau)},z_3^{(\tau)})$ 的

概率描述是假设有一个在3个变量上的概率分布 $p(z_1,z_2,z_3)$，并且在算法的第 τ 步已经选择了 $z_1^{(\tau)}$，$z_2^{(\tau)}$，$z_3^{(\tau)}$ 的值。

首先，将 $z_1^{(\tau)}$ 替换为新值 $z_1^{(\tau+1)}$，$z_1^{(\tau+1)}$ 来自于一个随机的抽样过程，其概率服从分布 $p(z_1 \mid z_2^{(\tau)}, z_3^{(\tau)})$。接下来，将 $z_2^{(\tau)}$ 替换为 $z_2^{(\tau+1)}$，$z_2^{(\tau+1)}$ 也来自于一个随机的抽样过程，其概率服从以下分布：$p(z_2 \mid z_1^{(\tau+1)}, z_3^{(\tau)})$。这里完成了一个流程，即先前的抽样 $z_1^{(\tau+1)}$ 可以作为下一步抽样的参数。同样的，对 z_3 进行类似的操作，其概率满足以下分布：$p(z_3 \mid z_1^{(\tau+1)}, z_2^{(\tau+1)})$，得到一组新的随机变量 $z_1^{(\tau+1)}$，$z_2^{(\tau+1)}$，$z_3^{(\tau+1)}$。

经过多次迭代后，样本与初始状态的分布无关。正是由于吉布斯抽样的特殊性质，下面陈述基于吉布斯抽样方法构建的大数据框架下的特征因子的筛选框架和落地实现流程。

2.2 基于吉布斯随机搜索算法对风险特征提取的实现路径流程

在大数据框架下，建模工作通常面对着高维度的稀疏数据，遍历高维度特征空间以提取出有效的特征进行建模是一项具有指数复杂度的工作，并不具备计算可行性，因此 Qian[24] 等提出了基于马尔科夫链蒙特卡洛模拟进行特征提取的方法，通过吉布斯抽样的方法将指数复杂度（NP 问题）的特征提取问题转化为在给定的（样本）误差容忍度（比如，5%的误差）下转化为多项式复杂度的问题，从而实现对高维特征空间的特征提取。基于 Qian[24] 等讨论的方法，按照如下步骤进行风险因子特征的提取：

第一步：建立初始化模型，构建初始特征集合。随机抽取一个特征子集 I_0 用于初步的模拟建模，将初始模型中系数不为 0 的特征记为 1，系数为 0 的特征记为 0，则有：

$$I_0=(0,1,1,\cdots,0)\in\{0,1\}^k。\tag{1}$$

第二步：基于 AIC[25]、BIC[5,26] 构建指标条件概率函数 $p(z)$（如式(2)、式(3) 所示）。

$$P_C(j_s=1\mid J_{-s})=\frac{P_C(j_s=1,J_{-s}\Rightarrow I_C)}{P_C(j_s=1,J_{-s}\Rightarrow I_C)+P_C(j_s=0,J_{-s}\Rightarrow I_C)},\tag{2}$$

式中，j_s 表示第 s 个特征；J_{-s} 表示除第 s 个特征之外的全部特征的组合；I_C 表示 J_{-s} 这一组合的确定值。然后，研究将分别基于 AIC（参见文献[25]）和 BIC 方法（参见文献[5]或文献[26]）构建两组条件概率分布函数，目的是在最后一步中比较两者的模型效果，条件概率分布函数可表示为：

$$\begin{cases}P_{AIC}(j_s=1\mid J_{-s})=\dfrac{exp(-AIC(j_s=1\mid J_{-s}))}{exp(-AIC(j_s=0\mid J_{-s}))+exp(-AIC(j_s=1\mid J_{-s}))}\\ P_{BIC}(j_s=1\mid J_{-s})=\dfrac{exp(-BIC(j_s=1\mid J_{-s}))}{exp(-BIC(j_s=0\mid J_{-s}))+exp(-BIC(j_s=1\mid J_{-s}))}\end{cases}。\tag{3}$$

第三步：进入抽样过程，共完成（通过下面推算出的）样本量为 400 次的模拟，计算每一个特征进入模型的频率，这个频率即表示该特征与被预测变量之间的关联显著性。对于各个特征的吉布斯抽样，由于假定随机分布满足伯努利过程，根据蒙特卡洛模拟的标准差公式，反映关联规则显著性的频率指标的标准差（用“$std(p)$”表示）如式(4) 所示：

$$std(p)=\sqrt{\frac{p(1-p)}{M}}<\sqrt{\frac{1}{4M}},\tag{4}$$

由式(4)可知，进行 400 次抽样可以保证样本关联显著性的误差小于 5%。这样就完成支持建立“风险特征因子筛选框架”相配套的在大数据框架下针对非结构性（风险）特征提取（筛选）的推断逻辑支持原理。

3 特征因子筛选框架方法在支持期货铜价格趋势预测特征因子刻画上的应用

利用建立的基于特征因子筛选框架方法，结合大宗商品期货产品铜和相关的真实市场和经济指标数据，讨论如何从众多（超过 126 个）与期货铜价格相关的指标（参见“附录 2”）中筛选出不超过 10 个与价格趋势变化的关联风险特征来进行预测的刻画如表 1、表 2、表 3 所示。这是过去很难想象和可以办到的事情。但是，利用大数据针对比较全面的数据进行有效地筛选，可以使大宗商品期货铜价格趋势变化（即价格变化的方向）方面的预测达到超过 90%的正确率。

3.1 大宗商品期货(铜)价格数据使用介绍

研究将以 2011 年 7 月至 2019 年 6 月之间沪铜期货指数合约价格（下文简称为“沪铜价格”）的每月累计涨跌幅度的方向为被预测变量，我们的工作是在今天（t 时间）预测期货铜价格在将来时间（大于 t）价

格变化的方向，因此，基于当下时间 t，我们预测的将来价格的变化只有两种情况：价格向上变化和价格向下变化（不失一般性，假定价格不变的概率为零）。

我们使用的描述“大宗商品期货产品铜”价格的指标为“当前市场中正在交易的所有同品种期货合约价格以成交量为权重的加权平均”。通常而言，剩余期限为 3 个月的期货合约持仓量最大（如图 1 所示），因此可以近似认为铜期货指数合约价格近似为 3 个月铜期货合约价格。

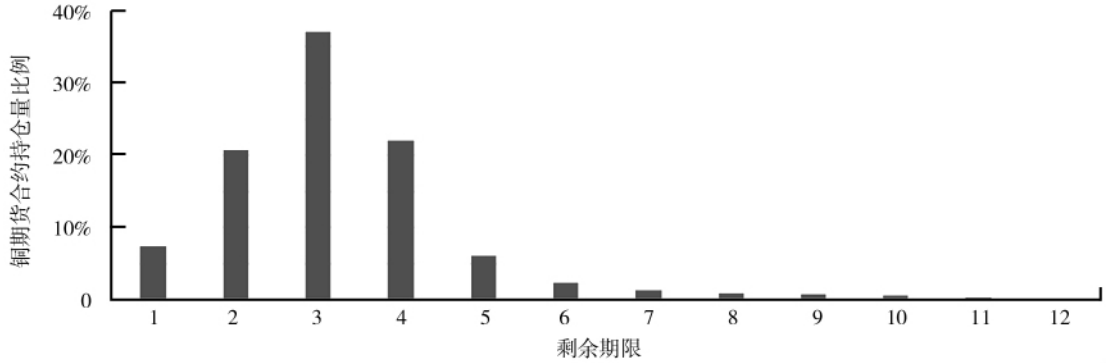

图 1　2011 年 7 月至 2019 年 6 月期间上期所铜合约日均持仓量比例

我们使用的其他解释变量则包括以下几个部分：商品期货指数合约行情数据、人民币兑美元中间价、沪深 300 指数及其行业子指数、宏观经济数据、ICSG（International Copper Study Group，国际铜研究组织）统计数据、精炼铜产量、出口量等数据。

由于预测是针对价格（将来）的变化趋势，自然需要考虑今天以及基于以前（针对时间序列的预测）时间点的可预测的特征因子，这是因为我们只能由变量滞后于被预测变量实时点的信息进行预测工作。“附录 2”中列出的 126（大）类初始特征因子作为最基本的出发点，考虑到因子滞后阶的四种情况：“1 个月，3 个月，6 个月和 12 个月”，这个构成了全部的初始因子，472 个作为最开始的备选解释变量（下面称为“初始特征”），下面汇报和讨论基于大数据特征提取方法获得的可以对期货（沪）铜价格变化的行情进行预测的具有强关联性的特征指标的表现情况（研究分析需要的数据全部来源于 Wind，参见 Wind 官方网址 https://www.wind.com.cn/）。

3.2　可以预测大宗商品期货铜价格变化趋势的关联特征因子和实证预测效果

基于上文所述的大数据特征提取方法以二分类（即价格向上变化和价格向下变化）Logistic 回归进行高度关联的特征提取。

为了研究沪铜价格趋势特征的变化情况，以每 5 年为一个时间窗口分别进行特征提取，因此分析的数据段分别为 3 个时间窗口：2011 年 7 月至 2017 年 6 月（简记为“11 年到 17 年”）；2012 年 7 月至 2018 年 6 月（简记为“12 年到 18 年”）和 2013 年 7 月至 2019 年 6 月（简记为“13 年到 19 年”）。

另外，在每个自然月内，若月内累计对数回报率大于 0，记为 1，表示当月沪铜行情为牛市；月累计对数回报率小于 0，记为 0，表示当月沪铜行情为熊市。

根据“附录 1”中 Odds Ratio 的概念，我们有下面基于吉布斯（Gibbs）随机搜索算法筛选出的关联风险特征因子的分类：①强关联（特征因子）：对应“特征比值比”小于 0.8 或大于 1.2 时；②一般关联（特征因子）：对应“特征比值比”介于 1.1 与 1.2 之间，或介于 0.8 与 0.9 之间；③弱关联（特征因子）：对应“特征比值比”大于 0.9 且小于 1.1 时。

首先考虑基于吉布斯（Gibbs）框架下，附录 2 中那些具有“关联显著性”表现的因子。“关联显著性”是指在吉布斯（Gibbs）随机抽样过程中，特征因子在模型中出现的概率（即公式（1）中定义的特征空间 I_0）。表 1 是具有“关联显著性”表现的前 14 个因子的汇总明细（基于“附录 2”中 126 个与铜价格相关的关联因子），这里“关联显著性”是指在研究建立的吉布斯（Gibbs）随机抽样过程中，特征因子在模型中出现的概率，即式（1）中定义的特征空间 I_0，表 1 中 ICSG 是该项指标的数据来源。

把表 1 中与铜价格变化相关的“强关联特征”因子分为三类。第一类为基础特征（编号 1～5），即反映沪铜的需求和供给的特征，同时也是最受沪铜交易者们关注的特征，经过大数据特征提取可见，产能缺口、精炼铜、铜材产量分别在不同年份中体现出与沪铜价格的强关联性。第二类为消费市场特征（编号 6～

9),即从产业链角度反映铜消费情况的特征,通过大数据特征提取发现家电行业(彩电、冰柜产量)、房地产面积(房地产竣工面积同比增长率)、基础设施建设(电网设施建设完成率)等特征同样是与沪铜价格趋势存在强关联的特征。由消费市场类特征的关联显著性可见,随时间推移家电产量增长率、房地产竣工面积增长率与沪铜价格趋势的关联显著性逐渐增强,而电网基本建设投资完成额同比增长率与沪铜价格趋势的关联显著性逐渐减弱。这一现象与我国当前电网建设逐渐趋向完善,国民消费升级的经济转型大趋势相吻合。第三类为宏观经济指标(编号 10～14),即反映宏观经济情况的常用指标。通过大数据特征提取后宏观指标并没有体现出很强的关联显著性,但是由于宏观经济指标具有对于经济整体状况的刻画能力,同时能够影响市场预期,因此在进一步对沪铜价格趋势的预测建模中仍将使用宏观经济指标作为特征。

表 1 与铜价格变化趋势相关的表现出"关联显著性"的 14 个因子名单

编号	价格关联特征(因子)	各时间窗内关联显著性/%		
		2011～2017 年	2012～2018 年	2013～2019 年
1	前 1 个月沪铜价格涨跌幅	100.00	98.50	99.75
2	前 1 个月 ICSG:期间库存变化	99.25	86.00	17.50
3	前 1 个月铜材产量同比增长率	89.25	53.25	17.00
4	前 1 个月精炼铜产量同比增长率	54.25	55.50	86.50
5	前 6 个月精炼铜产量(矿产)平均同比增长率	53.75	39.75	93.25
6	前 1 个月精炼铜(再生)同比增长率	64.25	32.00	96.00
7	前 1 个月彩电产量同比增长率	56.50	98.25	52.75
8	前 3 个月冷柜产量平均同比增长率	51.00	57.25	90.00
9	前 1 个月房地产竣工面积同比增长率	99.50	80.25	37.75
10	前 1 个月新增固定资产同比增长率	31.00	93.25	32.00
11	前 1 个月商务活动指数平均值	14.75	11.25	11.50
12	前 1 个月 PMI	13.00	12.50	9.75
13	前 1 个月 CPI 平均增长率	12.50	10.50	13.00
14	前 12 个月 GDP 累计值同比增长率	11.25	13.50	11.00

3.3 预测铜价格变化趋势的特征因子刻画

为了检验基于吉布斯随机搜索算法筛选出的对铜价格变化趋势可进行预测的高度关联的特征因子的有效性,采用上文提取得到的对筛选出的特征进行二分类逻辑回归建模筛选出的具有预测能力的关联方因子。

在检验过程中,使用 2011 年 7 月～2017 年 6 月的数据作为训练集,利用 2017 年 7 月～2019 年 6 月的数据为测试集检验提取得到的特征对样本外数据的预测效果的可靠性。

基于上面的讨论,在建模过程中采用二分类逻辑回归模型,将铜价格收涨的月份作为正例,记为 1;将收跌的月份作为负例,记为 0。同时,为了降低特征共线性对模型预测效果的影响,采用"$L1$"和"$L2$"两种正则化方法分别建模,基于模型对测试集数据的预测效果的好坏来验证建立的特征提取方法是否具备对沪铜价格变化趋势的预测能力。

正则化作为机器学习中常用的手段之一,本质是通过对拟合函数的损失行数添加一个正则化项,从而避免拟合函数出现过拟合的情况,并将拟合函数某些与结果不相关的自变量系数压缩为 0。$L1$ 正则化时,对应的惩罚项为 $L1$ 范数,即 $\Omega(\omega)=\|\omega\|_1=\Sigma_i|\omega_i|$;$L2$ 正则化时,对应的惩罚项为 $L2$ 范数,即 $\Omega(\omega)=\|\omega\|_2^2=\Sigma_i\omega_i^2$。

基于表 1 和表 2 的预测结果得出下面的基本结论:

结论 1:在 $L1$、$L2$ 两种正则化方法下建立的沪铜价格趋势预测模型对沪铜价格趋势预测的准确率分别为 95.83% 和 91.67%,两种模型均能较好地预测沪铜价格未来的变化趋势,如表 2 中的明细结果所示。

结论 2:从预测模型的模型系数来看,反映沪铜市场供需状态的基本特征(表 2 中编号 1～4)具有最强的解释能力;消费市场特征(表 2 中编号 6～9)能够在模型中对铜价格趋势的预测形成有效地补充;而宏观因子(表 2 中编号 10～14)的系数接近或等于 0,同时说明宏观因子与其他特征具有共线性,月度的宏观经济数据中的信息可以由基础特征和消费市场特征的线性组合所替代(至少在训练集样本数据内),即基

本特征与消费市场特征已经反映了宏观因子对沪铜期货未来价格的影响，其结果与数据特征提取中显示出的结果相吻合。

结论 3：通过比对基于 $L1$ 正则化方法和 $L2$ 正则化方法建立的模型结果，我们发现基于 $L1$ 正则方法的模型能够更好地对沪铜期货价格趋势进行预测，同时能够对指标进行进一步的提炼（如表 2 和表 3 中明细结果比较所示）。

综合起来，基于上面的 3 个结论并结合对应表 2、表 3 的结果，以及对应关联特征因子的 Odds Ratio 指标，有如下的结论：6 个高度关联的风险特征可以用来刻画期货铜价格变化趋势的预测（即对价格变化趋势的方向正确性达到 90％以上）。即，①前 1 个月沪铜价格涨跌幅；②前 1 个月 ICSG：期间库存变化；③前 1 个月精炼铜产量同比增长率；④前 1 个月精炼铜（再生）同比增长率；⑤前 1 个月房地产竣工面积同比增长率；⑥前 1 个月新增固定资产同比增长率。如果只是基于常规的计量分析的方法和手段，很难会发现“前 1 个月房地产竣工面积同比增长率”以及“前 1 个月新增固定资产同比增长率”会成为描述铜价格变化趋势的高度关联的特征刻画指标，这是大数据分析多维度信息融合的优点的体现。

表 2　沪铜价格趋势分析模型系数

编号	特征名称	$L1$ 正则化	$L2$ 正则化
1	前 1 个月沪铜价格涨跌幅	2.867 5	2.111 7
2	前 1 个月 ICSG：期间库存变化	－0.176 9	－0.214 3
3	前 1 个月铜材产量同比增长率	－0.008 0	－0.010 3
4	前 1 个月精炼铜产量同比增长率	－0.149 3	－0.103 1
5	前 6 个月精炼铜产量（矿产）平均同比增长率	0.000 0	0.035 3
6	前 1 个月精炼铜（再生）同比增长率	0.102 5	0.083 8
7	前 1 个月彩电产量同比增长率	－0.004 3	0.002 4
8	前 3 个月冷柜产量平均同比增长率	0.000 0	0.049 3
9	前 1 个月房地产竣工面积同比增长率	0.039 9	0.058 8
10	前 1 个月新增固定资产同比增长率	0.141 6	0.152 0
11	前 1 个月商务活动指数平均值	0.000 0	－0.293 1
12	前 1 个月 PMI	0.000 0	－0.407 9
13	前 1 个月 CPI 平均增长率	0.000 0	0.056 3
14	前 12 个月 GDP 累计值同比增长率	0.000 0	0.009 6
15	常数项	0.000 0	－0.001 2
	预测正确率（测试集数据）	95.83％	91.67％

表 3　沪铜价格趋势模型 Odds Ratio（基于 $L1$、$L2$ 正则化）

编号	特征名称	Odds Ratio（$L1$ 正则化）	Odds Ratio（$L2$ 正则化）
1	前 1 个月沪铜价格涨跌幅	17.593 0	8.262 3
2	前 1 个月 ICSG：期间库存变化	0.837 9	0.807 1
3	前 1 个月铜材产量同比增长率	0.992 0	0.989 8
4	前 1 个月精炼铜产量同比增长率	0.861 3	0.902 0
5	前 6 个月精炼铜产量（矿产）平均同比增长率	1.000 0	1.035 9
6	前 1 个月精炼铜（再生）同比增长率	1.107 9	1.087 4
7	前 1 个月彩电产量同比增长率	0.995 7	1.002 4
8	前 3 个月冷柜产量平均同比增长率	1.000 0	1.050 5
9	前 1 个月房地产竣工面积同比增长率	1.040 7	1.060 6
10	前 1 个月新增固定资产同比增长率	1.152 1	1.164 2
11	前 1 个月商务活动指数平均值	1.000 0	0.745 9
12	前 1 个月 PMI	1.000 0	0.665 0
13	前 1 个月 CPI 平均增长率	1.000 0	1.057 9
14	前 12 个月 GDP 累计值同比增长率	1.000 0	1.009 6
15	常数项	1.000 0	0.998 8
	预测正确率（测试集数据）	95.83％	91.67％

4 结论

研究的目的是建立对影响大宗商品期货价格变化趋势关联的(结构和非结构化)风险特征因子的提取框架和配套的推断逻辑原理。以金融科技中大数据概念为出发点,利用人工智能中的吉布斯随机搜索算法为工具,全面地陈述了如何提取高度关联大宗商品期货价格变化的风险特征因子的流程和配套的逻辑原理,即采用(在马尔科夫链蒙特卡洛(MCMC)框架下)人工智能中的吉布斯随机抽样算法,结合 OR 值作为验证标准(参见附录 2),实现从大量风险因子的数据中提取与大宗商品期货(铜)价格趋势变化相关的特征因子并进行分类,从而可用于构建支持期货价格趋势变化分析的特征指标。

研究的实证结果表明大数据特征提取算法能够有效地提取刻画沪铜指数合约价格趋势的特征,这些特征包含基础特征、消费市场特征和宏观经济指标多个维度,并利用这些特征为沪铜指数月度行情进行建模分析,最终实现较好的预测准确性。

特征挖掘的结果说明能够反映铜市场供需平衡状态的基础特征是对大宗市场进行预测分析的最有效特征,而消费市场特征能够在预测分析中起到有效补充。

基于 2011 年到 2019 年的真实数据,针对消费市场特征的变化进行分析,我们发现,用于能够刻画沪铜价格趋势变化的消费市场特征伴随着我国经济发展而变化,特别是随着基建设施的逐步完善,电网建设完成额与沪铜价格变化的关联性逐渐减弱;而随着消费升级的趋势,家电行业与沪铜价格的关联性逐渐增强。

我们的特征提取算法也表明:宏观经济指标与沪铜价格的变化关联性不强,但宏观经济指标具有对经济整体状况的刻画能力,同时能够影响市场预期,因此保留宏观经济指标作为特征因子是一个比较好的选择。

最后,希望研究建立的基于大数据框架下对刻画铜期货价格趋势变化(分析)的风险特征提取方法不只是理论上的创新,同时其结果可以用于业界实践指导铜期货的交易,风险管理和相关的资产投资业务的实践工作中。

参考文献:

[1] G R KRIPPNER.The financialization of the American economy[J].Socio-economic review,2005,3(2):173-208.

[2] K TANG,W XIONG.Index investment and the financialization of commodities[J].Financial analysts journal,2012,68(6):54-74.

[3] X Z YUAN,H Q WANG.The general dynamic risk assessment for the enterprise by the hologram approach in financial technology[J].International journal of financial engineering(IJFE),2019,6(1):1950001.

[4] M J BRENNAN,E S SCHWARTZ.Evaluating natural resource investments[J].Journal of business,1985,58(2):135-157.

[5] G SCHWARZ.Estimating the dimension of a model[J].The annals of statistics,1978,6(2),461-464.

[6] J CASASSUS,P COLLIN-DUFRESNE.Stochastic convenience yield implied from commodity futures and interest rates[J].The journal of finance,2005,60(5):2 283-2 331.

[7] 王苏生,王丽,李志超,等.基于卡尔曼滤波的期货价格仿射期限结构模型[J].系统工程学报,2010,25(3):346-353.

[8] E SCHWARTZ,J E SMITH.Short-term variations and long-term dynamics in commodity prices[J].Management science,2000,46(7),893-911.

[9] 韩立岩,尹力博.投机行为还是实际需求?——国际大宗商品价格影响因素的广义视角分析[J].经济研究,2012(12):84-97.

[10] 杨胜刚,陈帅立,王盾.中国黄金期货价格影响因素研究[J].财经理论与实践,2014,35(3):44-48.

[11] 朱晋.市场因素影响商品期货价格的多元模型分析[J].数量经济技术经济研究,2004,21(1):75-79.

[12] 张保银,陈俊.基于动态 VECM 的我国铜期货的价格发现功能研究[J].天津大学学报(社会科学版),2012,14(6):492-496.

[13] 董珊珊,冯芸.基于 FCVAR 模型研究 SHFE 和 LME 铜期货和现货市场价格发现功能[J].现代管理科学,2015(11):67-69.

[14] 黄健柏,刘凯,郭尧琦.沪铜期货市场价格发现的动态贡献——基于状态空间模型的实证研究[J].技术经济与管理研

究，2014(2)：67-72.
[15] J D HAMILTON，R SUSMEL.Autoregressive conditional heteroskedasticity and changes in regime[J].Journal of econometrics，1994，64(1-2)：307-333.
[16] 高辉，赵进文.期货价格收益率与波动性的实证研究——以中国上海与英国伦敦为例[J].财经问题研究，2007(2)：54-66.
[17] 张屹山，方毅，黄琨.中国期货市场功能及国际影响的实证研究[J].管理世界，2006(4)：36-42.
[18] 郭树华，王华，高祖博，等.金属期货市场价格联动及其波动关系研究——以 SHFE 和 LME 的铜铝为例[J].国际金融研究，2010(4)：79-88.
[19] 胡东滨，张展英.基于 DCC-GARCH 模型的金属期货市场与外汇、货币市场的动态相关性研究[J].数理统计与管理，2012(5)：150-158.
[20] Y D YUE，D C LIU，S XU.Price linkage between Chinese and international nonferrous metals commodity markets based on VAR-DCC-GARCH models[J].Transactions of nonferrous metals society of China，2015，25(3)：1 020-1 026.
[21] 李洁，杨莉.上海和伦敦金属期货市场价格联动性研究——以铜铝锌期货市场为例[J].价格理论与实践，2017(8)：100-103.
[22] 袁先智，狄岚，郭铁信，等.在大数据框架下基于 Gibbs 抽样的随机搜寻方法在金融风险特征提取中的应用[J].计量经济与金融学报，2020，1(1).
[23] A PAZ，S MORAN.Non deterministic polynomial optimization problems and their approximations[J].Theoretical computer science，1981，15(3)：251-277.
[24] G QIAN，C FIELD.Using MCMC for logistic regression model selection involving large number of candidate models[C]//In monte carlo and quasi-monte carlo methods 2000，Berlin Hedelberg：Springer-verlog，2002：460-474.
[25] H T AKAIKE.A new look at the statistical model identification[J].IEEE transactions on automatic control，1974，19(6)：716-723.
[26] S GERMAN.Gibbs distribution，and the bayesian restoration of images[J].IEEE Proc.pattern analysis and machine intelligence，1984，6：774-778.

附录 1　在大数据框架下支持关联特征提取的 Odds Ratio 指标描述

对于发生概率较小的风险事件(例如违约、发病等)而言，优势比(也叫比值比，英文标记为“Odds Ratio”)是相对危险度的精确估计值，接下来我们以吸烟和某种疾病的发病与否来解释 Odds Ratio 的意义。附表 1 解释吸烟与某种疾病的关系：其中吸烟的病例样本占总样本比例为 A，不吸烟的病例占总样本比例为 B；对照组(不发病的样本)中吸烟的样本占总样本比例为 C，不吸烟的样本占总样本比例为 D。

附表 1　Odds Ratio 案例

样本类型	吸烟者	不吸烟者
病例	A	B
对照	C	D

附表 1 中，病例样本中吸烟样本数量与不吸烟样本数量的比值为：

$$a=\frac{A}{B}, \tag{1}$$

病例中吸烟样本所占的比例越大，则该比值越大，也就意味着吸烟组所占的“优势”越大。但不能因这种“优势”较大就此简单认为吸烟提高了该疾病的发病风险。为了分析吸烟是否与该疾病的发病具有显著关联还需要对对照组进行分析，对照组中吸烟样本数量与不吸烟数量的比值为：

$$b=\frac{C}{D}, \tag{2}$$

两个比值的差异可以说明吸烟与该疾病的关联关系强弱，因此两个比值的比(即为 Odds Ratio) 如式(3)所示：

$$Odds\ Ratio=\frac{a}{b}=\frac{A*D}{B*C}, \tag{3}$$

Odds Ratio 越接近 1 则吸烟样本在病例组中所占的比例(或“优势”) 的差异越小，即吸烟与该疾病的关联关系越弱；*Odds Ratio* 越大或越接近 0 则说明，吸烟样本在病例组中所占的比例(或“优势”，下文也称为

Odds）的差异越大，即吸烟与该疾病的关联关系越强。

在 logistic 回归模型中，回归的因子通常为连续形变量（或是被处理称为连续型变量的），*Odds Ratio* 所反映的是因子变化 1 个单位带来的相对风险的变化。logistic 回归的 *Odds* 如式（4）所示：

$$\frac{p}{1-p}=exp(\beta_0+\beta_1 x_1), \tag{4}$$

式中，自变量为 x_1 的回归系数为 β_1。若自变量 x_1 增加 1 个单位后的 *Odds* 如式（5）所示：

$$\frac{q}{1-q}=exp(\beta_0+\beta_1(x_1+1)), \tag{5}$$

那么式（4）与式（5）的比值为特征 x_1 的 *Odds Ratio*，如式（6）所示：

$$Odds\ Ratio=exp(\beta_1), \tag{6}$$

由式（6）可见，特征在 logistic 回归中的 *Odds Ratio* 同样是一个常数，它不受特征的绝对数值的影响。

附录 2

附表 2　初步关联特征表

编号	特征名称	各时间窗内关联显著性/%		
		2011～2017 年	2012～2018 年	2013～2019 年
1	前 1 个月沪铜价格涨跌幅	100.00	98.50	99.75
2	前 1 个月房地产竣工面积同比增长率	99.50	80.25	37.75
3	前 1 个月 ICSG:期间库存变化	99.25	86.00	17.50
4	电网基本建设投资完成额:累计同比	96.25	94.25	33.25
5	前 1 个月铜材产量同比增长率	89.25	53.25	17.00
6	产量:精炼铜(铜):当月同比	79.75	58.50	79.75
7	前 1 个月精炼铜(再生)同比增长率	64.25	32.00	96.00
8	前 1 个月冷柜产量同比增长率	56.50	30.75	33.25
9	前 1 个月彩电产量同比增长率	56.50	98.25	52.75
10	前 1 个月精炼铜产量同比增长率	54.25	55.50	86.50
11	前 6 个月精炼铜产量(矿产)平均同比增长率	53.75	39.75	93.25
12	前 3 个月冷柜产量平均同比增长率	51.00	57.25	90.00
13	前 1 个月汽车产量同比增长率	46.75	19.50	27.00
14	前 3 个月彩电产量同比增长率	38.50	18.25	96.00
15	前 1 个月铜材库存同比增长率	32.50	23.25	35.50
16	前 1 个月新增固定资产同比增长率	31.00	93.25	32.00
17	出口数量:精炼铜:当月值	25.00	57.00	33.75
18	中央项目固定资产投资完成额_累计增长	25.00	28.50	24.75
19	产量:精炼铜(铜):矿产:当月同比	24.25	35.50	75.50
20	房屋施工面积_累计增长	21.50	69.75	76.00
21	产量:空调:当月同比	19.75	16.50	25.75
22	ICSG:期间库存变化:当月值	19.25	77.50	57.25
23	人民币兑美元中间价涨跌幅	15.75	12.50	12.00
24	出口数量:未锻造的铜及铜材:当月值	15.50	19.25	14.75
25	锌合约	15.50	15.75	10.25
26	房地产开发投资额_累计增长	15.25	16.25	44.00
27	沪深 300 可选消费指数	15.25	12.75	10.25
28	苹果合约	14.75	14.00	13.50
29	玉米淀粉合约	14.75	10.50	11.75
30	ICSG:期末精铜库存:当月值	14.75	10.25	11.75
31	前 1 个月商务活动指数平均值	14.75	11.25	11.50
32	纸浆合约	14.50	14.50	12.25
33	纤维板合约	14.25	11.50	13.75

编号	特征名称	各时间窗内关联显著性/%		
		2011～2017 年	2012～2018 年	2013～2019 年
34	沪深 300 主要消费指数	14.25	15.00	12.25
35	豆粕合约	14.25	15.50	11.50
36	石油沥青合约	14.25	11.00	9.75
37	ICSG：全球精炼铜产能：当月值	14.00	10.50	13.50
38	产量：精炼铜（铜）：再生：当月值	14.00	15.50	10.25
39	沪深 300 医药卫生指数	13.75	11.75	15.00
40	RU 天然橡胶合约	13.75	12.25	9.00
41	产量：彩电：当月值	13.50	14.50	14.50
42	ZC 动力煤合约	13.50	13.50	12.50
43	ICSG：再生精炼铜产量：当月值	13.50	9.50	11.50
44	PM 普麦合约	13.50	12.50	11.25
45	进口数量：精炼铜：当月值	13.25	13.25	16.75
46	JM 焦煤合约	13.25	10.00	13.75
47	I 铁矿石合约	13.25	12.50	11.75
48	出口平均单价：未锻造的铜及铜材：当月值	13.25	10.50	11.75
49	BB 胶合板合约	13.25	11.00	11.50
50	RO 菜籽油合约	13.25	10.75	10.75
51	产量：精炼铜（铜）：矿产：当月值	13.25	10.50	10.75
52	月度 CPI	13.00	13.00	17.75
53	ICSG：原生精炼铜产量：当月值	13.00	12.50	11.50
54	前 1 个月 PMI	13.00	12.50	9.75
55	地方项目固定资产投资累计完成额同比增长率	12.75	14.75	21.25
56	铜材产量（当月值）	12.75	12.00	15.75
57	B 豆二合约	12.75	12.25	14.50
58	AU 黄金合约	12.75	14.50	14.00
59	CJ 红枣合约	12.75	11.75	13.75
60	TC 动力煤合约	12.75	12.50	13.00
61	沪深 300 原材料指数	12.75	14.75	12.50
62	产量：精炼铜（铜）：当月值	12.75	9.25	12.50
63	WT 硬白小麦合约	12.75	12.00	11.25
64	月度 GDP	12.75	11.75	10.50
65	沪深 300	12.75	13.00	10.25
66	PTA 合约	12.50	14.25	13.25
67	ICSG：全球矿山产能：当月值	12.50	13.00	13.00
68	RI 早籼稻合约	12.50	12.00	12.75
69	前 1 个月 CPI 平均增长率	12.50	10.50	13.00
70	ICSG：全球精炼铜产量（原生＋再生）：当月值	12.25	13.75	11.50
71	ICSG：精炼铜产能利用率：当月值	12.25	13.25	11.00
72	RM 菜籽粕合约	12.00	12.50	13.00
73	沪深 300 公用事业指数	12.00	11.25	13.00
74	LR 晚籼稻合约	12.00	12.25	11.75
75	JR 粳稻谷合约	12.00	13.00	11.00
76	AG 白银合约	12.00	15.00	10.00
77	产量：冷柜：当月值	11.75	12.25	16.50
78	P 棕榈油合约	11.75	11.75	14.75
79	RB 螺纹钢合约	11.75	12.25	14.75
80	进口数量：未锻造的铜及铜材：当月值	11.75	12.75	14.25

编号	特征名称	各时间窗内关联显著性/%		
		2011～2017 年	2012～2018 年	2013～2019 年
81	沪深 300 信息技术指数	11.75	11.25	13.50
82	WH 强麦合约	11.75	10.75	13.25
83	AL 铝合约	11.75	10.25	12.75
84	V 聚氯乙烯合约	11.50	10.00	13.50
85	CY 棉纱合约	11.50	12.75	13.25
86	NI 镍合约	11.50	16.00	12.75
87	FU 燃料油合约	11.50	10.50	11.75
88	SC 原油合约	11.50	11.25	11.00
89	FG 玻璃合约	11.50	12.00	10.50
90	MA 甲醇合约	11.50	15.00	10.00
91	产量:汽车:当月值	11.25	13.25	17.00
92	RS 油菜籽合约	11.25	11.25	13.50
93	沪深 300 工业指数	11.25	14.25	13.00
94	PB 铅合约	11.25	10.75	12.25
95	SN 锡合约	11.25	13.25	12.00
96	ICSG:全球精炼铜消费量:当月值	11.25	13.25	11.50
97	JD 鸡蛋合约	11.25	13.25	11.25
98	CF 棉花合约	11.25	11.50	10.25
99	进口平均单价:未锻造的铜及铜材:当月值	11.25	13.25	8.25
100	前 12 个月 GDP 累计值同比增长率	11.25	13.50	11.00
101	产量:空调:当月值	11.00	8.50	16.00
102	HC 热轧卷板合约	11.00	13.75	14.00
103	电网基本建设投资完成额:累计值	11.00	11.25	14.00
104	EG 乙二醇合约	11.00	8.00	13.00
105	沪深 300 电信业务指数	11.00	14.50	12.25
106	SM 锰硅合约	11.00	10.50	10.75
107	PP 聚丙烯合约	11.00	12.00	10.25
108	商务指数	10.75	16.50	39.50
109	国内生产总值同比(累计值)	10.75	13.00	12.75
110	Y 豆油合约	10.75	11.00	12.25
111	ICSG:矿山产能利用率:当月值	10.75	9.50	10.75
112	SF 硅铁合约	10.50	12.00	16.00
113	CPI 同比(与去年同期相比)增长率(用 cpi_yoy 表示)	10.50	14.50	15.00
114	销量:汽车:当月值	10.50	11.00	12.75
115	C 玉米合约	10.50	13.25	12.25
116	沪深 300 金融地产指数	10.50	11.50	11.00
117	L 聚乙烯合约	10.50	9.75	10.25
118	SR 白糖合约	10.25	11.50	13.00
119	A 豆一合约	10.25	13.50	11.50
120	PMI	10.00	13.50	19.00
121	ER 早籼稻合约	10.00	12.00	13.75
122	沪深 300 能源指数	10.00	11.75	13.50
123	OI 菜籽油合约	10.00	8.75	12.00
124	WR 线材合约	9.75	10.00	12.00
125	GN 绿豆合约	9.25	13.25	10.75
126	J 焦炭合约	9.00	12.50	12.00

The Framework of Extract for Related Risk Factors by Using AI Algorithms and Applications to the Forecast of Trend for Commodity Futures Prices in Practice

YUAN George[1,2,3], ZHOU Yunpeng[3*],
LIU Haiyang[3*], YAN Chengxing[3*],
QIAN Guoqi[4], QIAN Xiaosong[5],
WANG Donghua[6], LI Zhiyong[7],
LI David[8], LIN Jianwu[9],
SHEN Sicheng[3], ZENG Tu[3]

(1.Business School, Chengdu University, Chengdu 610106 China;
2.School of Financial Technology, Shanghai Lixin University of Accounting and Finance, Shanghai 201620 China;
3.BBD Technology Co., Ltd.(BBD), No.966 Tianfu Avenue, Chengdu 610093, China;
4.School of Maths & Stats, The University of Melbourne, Melbourne VIC3010, Australia;
5.Center for Financial Engineering, Soochow University, Soochow 215006, China;
6.Business School, East China University of Science and Technology, Shanghai 200237 China;
7.School of Finance, Southwest Univ.of Finance and economics, Chengdu 611137 China;
8.Shanghai Advanced Institute of Finance, Shanghai 200030 China;
9.Tsinghua Shenzhen International Graduate School, Shenzhen 518057 China)

Abstract: Based on the available macro and micro factors, this paper creates a massive pool of original data of both structured and unstructured factors and analyzes the effects of the factors from the pool on the change of the price of the relevant commodity futures. The method of Gibbs sampling of logistic regression candidate models is used for effective and scalable screening of all features, resulting in identifying those macro and micro-factors, with importance weighting measures, that influence the commodity price change. The empirical results show that the Gibbs sampling induced big data feature extraction algorithm can effectively extract the features related to the price trend of the Shanghai copper index contract. Further analysis of the associations between the trend of price changes and the identified effecting features reveals an important and sensible explanation of the social and business environment of the feature mapping of the futures association. The paper points out that our method of risk features extraction based in the big data framework is not only an innovation in theory for depicting the copper futures price trend, it also has technical innovations providing effective guidance in future copper trading in industrial practice.

Key words: big data; Gibbs sampling; stochastic search; feature extraction; related parties; forecast of price trend