中国管理科学
Chinese Journal of Management Science
ISSN 1003-207X,CN 11-2835/G3

《中国管理科学》网络首发论文

题目：财务欺诈风险特征筛选框架的建立和应用
作者：袁先智，周云鹏，严诚幸，刘海洋，钱国骐，王帆，韦立坚，李志勇，李波，李祥林，曾途
DOI：10.16381/j.cnki.issn1003-207x.2020.2201
网络首发日期：2021-05-11
引用格式：袁先智，周云鹏，严诚幸，刘海洋，钱国骐，王帆，韦立坚，李志勇，李波，李祥林，曾途．财务欺诈风险特征筛选框架的建立和应用．中国管理科学.
https://doi.org/10.16381/j.cnki.issn1003-207x.2020.2201

网络首发时间：2021-05-11 09:35:04
网络首发地址：https://kns.cnki.net/kcms/detail/11.2835.g3.20210510.1731.003.html

DOI：10.16381/j.cnki.issn1003-207x.2020.2201

财务欺诈风险特征筛选框架的建立和应用

袁先智[1,2,3]，周云鹏[3]，严诚幸[3]，刘海洋[3]，钱国骐[4]，王帆[2]，韦立坚[2]，
李志勇[5]，李波[6]，李祥林[7]，曾途[3]

（1.成都大学商学院，四川成都 610106；
2. 中山大学管理学院，广东广州 510275；
3. 成都数联铭品科技有限公司（BBD），四川成都 610093；
4.墨尔本大学数学与统计学院，澳大利亚墨尔本 VIC3010；
5.西南财经大学金融学院，四川成都 611137；
6.重庆理工大学理学院，重庆 400054；
7.上海高级金融学院，上海 200030）

摘　要：本文从金融科技大数据出发，以人工智能的吉布斯随机搜索（Gibbs Sampling）算法为工具，在大数据框架下建立了针对公司财务欺诈风险的特征因子筛选的一般处理方法与特征提取推断原理，并结合上市公司的财务报表数据进行实证分析，结合从 2017 年 1 月到 2018 年 12 月证监会对上市公司财务报表信息披露违规的数据样本，筛选出刻画财务欺诈的特征因子并进行了验证测试，支持财务欺诈的识别。本文提出的框架和模型方法可以加强和提升对上市公司财务欺诈风险的识别能力，并实现对公司财务在欺诈方面的探测与预测（Detecting and Predicting）功能。

关键词： 大数据；吉布斯随机搜索（Gibbs Sampling）抽样；随机搜索算法；SAS99；财务欺诈风险；舞弊三角理论；特征提取推断原理

中图分类号：F803.9 0212.2 TP182　**文献标识码：**A

1　引言

随着金融科技的发展，大数据的思维框架和机器学习方法的快速发展为财务欺诈识别提出了新的解决思路。在量化投资逐渐成为主流的今天，Beneish[1]提出的 *M-Score* 方法为基于量化分析进行财务欺诈风险建模提出了初步的方法。但在更一般的财务欺诈风险识别与管理的领域需要量化分析工具为服务于不同目的的尽职调查工作提供指引，因此也对财务欺诈风险的模型计量与刻画方法提出了更高的要求。为了服务于各种不同应用场景下的需求，在大数据的背景下需要一种能够根据不同的目的具体需求而进行特征提取的方法，从而更有效地支持服务于不同目标的财务欺诈识别与风险管理等应用场景。

对企业的分析和评估，最具代表性的方法是 Palepu 等[2]提出的从战略，会计，财务，前景四个方面对公司进行全面分析，即有名的哈佛分析框架。哈佛分析框架的核心思想是基于多维度的融合分析方法，即对公司进行分析不能孤立从财务或其它单一的方面进行，而是应该根据各方面信息来得出综合的分析结论，同时它也强调了会计报表质量对于公司评估的重要性。美国注册会计师协会(AICPA)[3]在其标准 SAS99（和 SAS82）《财务报表审计中对欺诈的考虑》中把财务欺诈定义为“在财务报表中蓄意错报，漏报或泄露以欺骗财务报表使用者”。纵观全球资本市场，上市公司财务欺诈都是资本市场中不可忽视的一类事件。一方面由于财务欺诈事件会给投资者带来巨大的损失，另一方面是上市公司通常都是具有一定影响力的（集团或实体）公司，这些公司的财务欺诈所引发的连锁反应可能演化成为系统风险（例如安然事件）。Niu 等[4]对中国市场的实证研究发现公司欺诈行为会对投资者行为方式产生重大的影响，使得投资者变得更加保守，从而对资本市场造成伤害。Healy 和 Palepu [5]从信息不对称的角度对公司信息披露进行了研究，提出了对公司披露信息进行分析研究的方法框架，并对各类信息披露监管法规，披露渠道方面的研究进行了梳

基金项目：国家自然科学基金资助项目（U1811462）；
国家自然科学基金资助项目（71971031）
作者简介：袁先智(1965-)，男(汉族)，重庆人，成都大学商学院特聘教授，博士，研究方向：金融科技，金融工程，风险管理，共识博弈论，区块链与共识经济。
通讯作者简介：周云鹏(1993-)，男(汉族)，云南昆明人，成都数联铭品科技有限公司金融科技分析师，硕士，研究方向：金融工程，E-mail：aviyp@outlook.com.

理和总结，此外他们的研究还发现了许多在业界实践中尚未得到落地实施和需要解决的基础性问题。Defond 和 Zhang[6]从衡量审计质量的角度进行了研究，提出了衡量审计质量的方法框架，Donovan 等[7]在他们的基础上对于审计质量评估的方法进行了进一步的研究。而 Yang 和 Lee[8]从法务会计的角度对企业欺诈风险管理的角度进行研究，提出了以平衡评分卡为基础的评估方法，为公司治理，反欺诈等提供了决策工具。Vanhoeyveld 等 [9]对税务的层面的欺诈行为进行了研究，通过无监督异常检测的方法对增值税欺诈的问题提出了解决方案。Nurhayati[10]和 Goode 和 Lacey[11]也讨论公司内控管理制度与财务欺诈风险之间的关联关系。Beasley[12]则从董事会和监事会成员人数的角度对引发财务欺诈风险进行了研究。众多学者们从不同角度进行了财务欺诈相关的分析，但公司财务欺诈风险的分析涉及到公司的经营、管理、财务、法务、公司治理、信息披露和监管等方方面面，因此在实务中对公司的欺诈风险进行甄别时需要根据不同的目标特征表现进行繁杂的专业分析和配套的尽职调查。

本文在大数据的框架下针对公司的财务欺诈风险进行特征刻画，综合企业财务信息，会计信息，生态信息来全面地构建上市公司财务欺诈风险模型。本文通过马尔科夫链蒙特卡洛 (MCMC) 框架下的吉布斯随机搜索方法（Gibbs Sampling）对特征子集进行随机抽样，在给定样本误差容忍度的条件下(假定在 5%的样本误差)，提出了基于上市公司财务报表数据分析的财务欺诈特征提取方法和对应的特征提取推断原理，解决了由于考虑财务报表勾稽关系而产生的维数灾难问题，从财务报表数据以及各财务数据的勾稽关系中提取出 8 个与财务欺诈高度关联的风险特征因子，他们能够有效地刻画上市公司的财务欺诈风险。同时结合这 8 个特征指标的比值比（Odds Ratio）[13]和其会计含义进行分析发现，与上市公司财务欺诈风险具有高度关联性的特征通常与公司的会计政策选择、公司治理等因素也具有高度的关联性，这一点与传统的财务报表分析框架相符。实证数据结果显示本文方法能够比较有效的对刻画财务欺诈的特征进行提取，从而在大数据框架下为审计，合规，投资分析等场景下的尽职调查和风险分析提供技术支持。

2 公司财务欺诈风险关联的特征与特征筛选方法陈述

财务分析是进行财务欺诈识别的工作中一个极其重要的环节[1-18]，但在真实的场景中财务欺诈活动具有高度的动态性和不确定性，需要综合财务与非财务的因素进行综合分析才能得出最终的结论。针对这种动态性和不确定性，本文认为在大数据的框架下应该通过对财务欺诈的特征刻画来模拟一家公司的财务欺诈风险，而不是定性判断一家公司是否在其真实的业务活动和信息披露中存在财务欺诈。但为了实现大数据框架下的财务欺诈风险特征刻画，我们首先需要对传统的分析方法中构建相对一般化的初始特征，或者说寻找通过基于人工智能大数据分析的方法来构造初始特征集合。同时，考虑到初始特征集合可能包含的关联因子比较多，各个特征之间还会存在交互效应使得特征筛选工作面对典型的 *NP* 问题(参见 Paz 和 Moran[14])，为了克服 *NP* 问题，本文通过梳理关联规则学习的方法，建立在大数据框架下针对上市公司财务欺诈风险特征筛选的基于人工智能的吉布斯随机抽样搜索（Gibbs Sampling Search）算法来完成针对刻画公司财务欺诈风险的特征筛选。

2.1 上市公司财务欺诈风险特征

上市公司进行财务欺诈活动的原因可能是多种多样的，其表现形式和实施手段也是随着社会、经济的发展进程而持续演化的，因此财务欺诈活动具有高度的不确定性和动态性。陈竞辉和罗宾臣[15]针对亚洲上市公财务欺诈案例进行研究后指出各类不同的财务欺诈案例都显示出了公司治理的不足是财务欺诈的重要特征，但是由于行业特点、监管要求等各种因素的变化，在每一家公司中的公司治理问题也会以完全不同的形式表现出来。叶金福[16]基于国内财务欺诈样本进行研究也得出了类似的结论，指出复杂的股权结构、资金流动缺乏痕迹、业务环节难以验证、高风险的会计政策等是财务欺诈的重要风险因子，同时，他的研究还指出由于经济活动的组织形式随着社会发展的步伐也在不断演化，而且在不同行业的财务欺诈也会有不同的表现形式和特征，而很多行业的经营特征难以在短时间内形成一般性的经验和结论。刘姝威[17]则系统阐述了财务欺诈识别需

要从财务分析、基本面分析 (包含宏观经济、行业特征、公司治理、管理能力、经营特征等多方面因素)、现场调查综合分析后才能得出结论，同时指出现场调查应该是判断财务欺诈的核心环节。结合真实的案例来看，对于公司业务线相对清晰的公司如银广夏、蓝田股份、康得新、雏鹰农牧等，分析人员可以对公司的财务数据、业务数据、资产凭证等信息进行综合分析来定位财务欺诈的原因，并查找相关证据。但是对于多元化经营的集团 (如德隆) 而言，错综复杂的集团生态网络关系很可能掩盖其中的利益输送等关系，对于其中是否存在财务欺诈的问题而言同样难以定性。王昱和杨珊珊[19]在研究上市公司财务困境中对财务数据指标体系也进行了分类研究，发现资产规模、资本结构、偿债能力等 21 个财务比率可以建立财务预测指标体系。洪文洲等[20]也在 2004-2013 年的时间段中选择了 44 家财务欺诈舞弊的公司和 44 家正常经营的上市公司进行了财务舞弊指标（27 个）的对比验证。周利国等[21]则将公司的财务数据（利息保障倍数，总资产周转率等）作为研究企业集团信用风险传染效应的微观协变量，结合宏观协变量确定公司的违约距离。以上的文章尽管研究的焦点不同，但是涉及到公司是否存在违约、欺诈以及财务困境，都离不开对公司底层财务数据的分析。

基于上述的研究和商务、财务、会计报表之间的关系，可以知道所有的财务欺诈活动都会在财务报表及关联方信息中留下线索和痕迹，这就使得通过财务大数据的全息画像方法（也称为 Hologram，参见 Yuan 和 Wang[17]）对上市公司的财务欺诈进行多维度的刻画成为可能。面对财务欺诈的高度不确定性和动态性，在大数据框架下进行特征刻画时不应该从定性判断的角度入手，而是应该从风险计量的角度来模拟上市公司存在财务欺诈的风险。这种风险特征刻画的思想可以在更高效地规避风险的同时节省大量用于投资研究的时间，在金融科技快速发展、量化分析逐渐占据主流的今天更能满足量化投资、信用评级等现实应用场景中的实际需求。

2.2 基于财务分析的财务欺诈识别方法陈述

财务分析是对甄别财务欺诈必不可少的一个步骤，而在大数据框架下，传统的财务分析方法同样能够为财务欺诈的特征提取工作提供构建初始特征池的基本思路。因此本文工作的第一步需要对目前已经在学术界和实务界中最常用的财务分析方法进行梳理，形成构造初始特征池的指标。陈竞辉和罗宾臣[15]按照虚增利润，夸大绩效，虚增资产，虚减负债，伪造现金流，五个维度进行财务分析，并详细介绍了如何在各个模块中应用财务比率进行分析。叶金福[16]主要针对毛利率和现金流两个角度提出分析方法，针对上市公司的主营业务活动中的会计方法，应收账款，存货等问题的分析方法进行了详细阐述，着重介绍了对公司的收入和费用结构的分析方法，并结合财务报表的勾稽关系指出在存在盈利操纵的情况下会产生哪些资产和负债项目的变化。刘姝威[17]提出的分析方法则是按照静态分析，趋势分析，同业比较的三个模块来进行的，其中静态分析即是对定时期或一个时间点上的财务数据和财务指标进行分析;趋势分析即是对不同时期的财务数据和财务指标进行分析；同业比较则是将一家公司的财务数据和财务指标与同行业的企业相比较。基于他们的分析方法，本文将会基于以下两个基本的出发点对财务数据进行分析和处理来构造初始特征池：

（1）财务分析时应该全面地涵盖对收入、费用、资产、负债、现金流的全方位分析;

（2）分析时应该综合考虑各个不同时期的财务数据，并且以同行业的其他公司的财务数据作为参照。

2.3 特征提取方法简述

在面对海量数据时，通过算法自动发现的特征之间的关联关系即为特征提取。在统计学中，相关性检验能够反映特征之间是否存在线性相关性 (例如使用皮尔逊相关系数)，但是在大数据框架下大量的特征之间的关联性都是非线性的，难以通过相关性来描述。另一方面，在面对高维特征空间的时候，很难通过变量之间的两两线性相关关系找到最适合用于建模的特征子集，而遍历特征空间则会面临典型的 NP 问题，算法会因为指数级的算法复杂度而在面对高维特征空间时失去计算可行性，而正则化方法在特征空间维数接近甚至超过观测样本数量时极有可能无法收敛。综上所述，在大数据的背景下，对高维的特征空

间进行特征提取时难以避免两个难题，一是由于特征之间 (包括特征与响应变量) 的关联关系不在只是线性的相关关系;二是特征空间维度过高而观测样本数量有限的矛盾。为了解决上述两个难题，本文采用 Logistic 回归的方法来刻画上市公司的财务欺诈风险，并借鉴关联规则学习算法解决特征维度过高的思想，在基于马尔科夫链蒙特卡洛模拟 (MCMC) 框架下的吉布斯随机搜索（Gibbs Sampling）算法在观测样本量有限的条件下降低计算复杂度。

关联规则学习是显示数据中特征之间关联关系的技术，目前被广泛的应用于零售、金融、Web 用户行为分析等领域。例如，通过对用户的网页浏览数据进行分析可能会发现常在购物网站搜索剃须刀的用户同时还可能需要搜索什么别的商品，从而准确地向用户推送相关产品网页链接。由于这些应用场景中常会面对大于观测样本数量的特征数量（即上文提到的特征维度过高而观测样本数量有限的矛盾），因此关联规则学习中通常都会针对这一问题提出解决方案。

目前最具有代表性的关联规则挖掘算法包括 Apriori[22]和基于吉布斯随机抽样(Gibbs Sampling)的特征挖掘算法等。Apriori 算法由 Agrawal 和 Srikant[22]提出，基本思想是利用项集支持度的某种性质避免了穷举所有候选项集 (关联规则学习相关概念解释参见欧高炎等[23])，从而解决了遍历特征空间的 NP 问题[14]。除了 Apriori 方法外，Qian 和 Field[24]提出了基于吉布斯随机抽样（Gibbs Sampling）的特征挖掘算法，其基本思想是利用吉布斯随机抽样在复杂采样过程中不易造成偏差的特性从复杂的多元概率分布中产生随机向量，实现对特征空间进行随机抽样的同时保证所抽取随机样本能够保持特征的原始信息，从而将 NP 问题转化为多项式级复杂度的问题，这就解决了高维特征空间中的关联规则学习问题[25]。

在实证研究中 Qian et al.[26]基于关联规则的置信度构造转移概率，针对 229 个病例 (其中 39 个为乳腺癌) 病例的基因片段数据 (包含 366 个基因编码区) 进行关联规则学习得到了优于 Apriori 的效果，在设定最小支持度为 0.2，最小置信度为 1 的条件下，通过 6000 次模拟获得了 35 条重要的关联规则，但在设定相同的最小支持度和置信度的条件下 Apriori 算法并不能提取出有效的关联规则。

吉布斯随机抽样作为一种简单有效的马尔科夫链蒙特卡洛模拟 (MCMC) 方法在学术界和实务界都有广泛的应用。Glasserman[27]应用 MCMC 方法在金融领域进行了大量的应用研究。Narisetty 等[28]讨论了一种支持模型选择的可伸缩 Gibbs Sampling 算法。袁先智等[29]则应用 Gibbs Sampling 对刻画大宗商品价格趋势的关联特征进行了研究，并基于蒙特卡洛模拟的性质建立了如何设定模拟次数来控制样本量与筛选的关联特征因子带来的误差显著性关系。

2.4 马尔科夫链蒙特卡洛下的吉布斯抽样方法

正如上面 2.3 节所述，大数据中的许多规划问题难以得到精确的答案，利用蒙特卡洛模拟框架下的吉布斯随机搜索算法，可以使得面对海量数据并在一定的误差容忍度下，花费合理的计算资源完成对问题的求解，得到一个近似解，故本文建立使用基于吉布斯随机搜索算法的一般框架来实现对公司财务欺诈风险特征的筛选处理。

首先，假定财务欺诈风险特征因子服从伯努利分布，对特征因子形成的特征空间进行初始化，并进行随机抽样，将特征根据系数是否为 0 进行分类，不为 0 的记为 1，为 0 的记为 0，可得：

$$A_0 = (0, 1, 1, \dots, 0) \in \{0,1\}^m \quad (1)$$

其中m表示在初始化的特征空间中的特征个数，A_0表示在初始化的特征空间中的一个子集。

然后，通过 BIC[30] (Bayesian Information Criterions)构建支持随机抽样的标准，并构建出特征的分布函数，得：

$$P_{BIC}(i_n = 1|I_{-n}) = \frac{exp(-BIC(i_n=1|I_{-n}))}{exp(-BIC(i_n=0|I_{-n}))+exp(-BIC(i_n=1|I_{-n}))} \quad (2)$$

其中$P_{BIC}(i)$表示指标转移概率函数，i_n表示第 n 个特征，I_{-n}是除i_n外的其他特征集合，在初始化的特征空间中的特征个数，I_0表示在初始化的特征空间中的一个子集，利用该公式来保证特征子集向拟合度更高的方向转移，使得最终的财务欺诈指标的显著性得以显现。

进而，确定样本的抽样次数。确定抽样次数是为了降低计算复杂度，让最终指标显著性的结果在可容忍误差范围内得以实现。此时，我们需要设定误差范围，为了保证财务欺诈指标的显著性，样本量误差通常建议为不超过 5%，其对应的

公式如下：

$$Std(p)=\sqrt{\frac{p(1-p)}{M}}<\sqrt{\frac{1}{4M}} \quad (3)$$

当以 *2-sigma*（即，2 *Std(p)*）准则来控制模拟误差在 5%以内，通过公式（3）可求得抽样次数M大于等于 400 次。该抽样次数可以起到降低计算复杂度并保证特征显著性的效果。

最后，进行不小于 400 次的吉布斯抽样，得到特征指标的组合$(I^{(1)},I^{(2)},\dots I^{(M)})$，利用特征出现的次数与抽样总次数的比值，求得特征出现的频率，根据频率的高低分析特征对模型结果的影响。通过受试者工作特征（ROC）(即 Receiver Operating Characteristic）下方面积（AUC）(即 Area Under the CurveROC）作为模型的评价标准来衡量最终得到的特征指标的显著性（敏感性）。

3 在大数据框架下基于吉布斯随机搜索方法对上市公司财务欺诈特征的提取

基于前面的介绍，本节的重点是以上市公司全体的财务报表数据为基础，利用人工智能的吉布斯随机搜索（Gibbs Sampling）算法为工具建立在大数据框架下筛选描述公司财务欺诈风险的特征因子的一般处理方法与实施原理，并结合真实的非结构化坏样本，形成基于上市公司全样本的实证分析。特别地，根据 2018 年 A 股上市公司财务报表构建支持财务欺诈风险的初始特征因子集合为出发点，结合从 2017 年 1 月到 2018 年 12 月中国证监会对上市公司财务报表信息披露违规的数据样本为基础寻找可描述“财务欺诈风险”的特征进行监督学习，然后筛选出可刻画财务欺诈的特征因子：即，通过解决由于财务报表勾稽关系而产生的维数灾难问题，从财务报表数据中提取出为数不多的（在本文为 8 个）关联的特征因子来刻画上市公司的财务欺诈风险。同时，结合刻画财务欺诈特征指标的比值比（Odds Ratio）作为验证标准，发现这些指标呈现出与财务欺诈所在公司的会计政策选择、公司治理等因素具有高度关联性的特点。

3.1 构建刻画财务欺诈行为的初始特征集合

在传统财务分析方法的启发下，本文将根据 2.2 的论述来构建初始特征池，将特征构造方法分为静态分析、动态 (趋势) 分析、同业比较三类，并在此基础上，实现对财务报表钩稽关系的刻画，从而支持筛选出可以刻画财务欺诈风险的特征指标。

（1）静态分析

静态分析即为对财务报表结构进行分析，考虑资产负债结构、收入费用结构、现金流与收入费用、资产负债的钩稽关系等因素。在进行静态分析建模时我们选择采用百分比报表的方法对财务数据进行预处理。采用百分比报表既能反映被评估公司的资产负债结构，又能实现了对被评估公司财务数据的归一化。基于百分比报表的方法进行归一化后少部分公司的财务数据仍然可能存在离群值 (如投资收益显著高于营业收入)，因此还需要对异常值进行处理后才能参与建模。

（2）动态 (趋势) 分析

动态 (趋势)分析用于捕捉公司的财务数据发生的异常变化。基于会计报表的勾稽关系，舞弊活动粉饰了一部分财务科目的同时通常会使另外一些财务科目发生异常变化。另一方面，考虑到虚增资产、虚构利润、虚构现金流等财务舞弊活动是一项非常庞大的系统工程，通常需要多部门甚至多公司主体配合，当舞弊手段难以获得商务、税务等各方面活动进行支撑时通常可能会产生较大额度的资产减计，从而在财务报表数据中发生异常的变化。为了捕捉上述两种异常变化，我们采用各个财务报表科目最近一个年度的变化量为基础，再使用过去四个年度的相同科目的变化量形成的增长率指标进行分析，从而得到衡量当前一个年度被评估公司的财务指标变化的异常程度。

（3）同业比较

即是基于同行业的平均表现来提取财务欺诈的特征，在具体方法为将同行业的样本作为一个集合，再以 Z-score 标准化的方式对同行业内的数据进行标准化，使得处理后的数据具有固定的均值和标准差。值得注意的是我们并未假设任何财务指标在同行业内应该服从正态分布，而是通过单个样本与同行业的平均值之间的差异来定义其与同行业其他样本之间的差异，再通过正态分布累计概率函数进行归一化处理。

（4）对财务报表钩稽关系的刻画

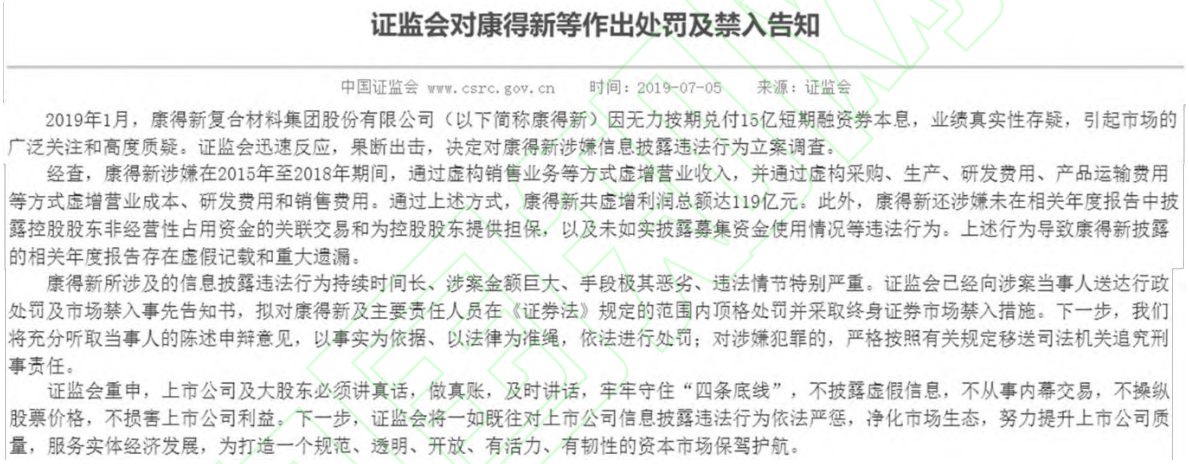
考虑财务报表之间的勾稽关系，本文不仅需要考虑到财务数据本身可能出现的异常，也应该考虑各个财务报表科目之间的比率关系。

本文以（非金融行业）在 A 股上市的实体公

司在 2018 年的年度报告财务数据为基本，从上市公司筛选出3459个样本公司的三大财务报表中的财报科目作为基础数据，按照静态分析、动态 (趋势) 分析、同业比较三个模块的数据处理方法构造初始特征池（近 200 个财务报表科目的初始因子），然后利用吉布斯随机搜索针对特征提取的方法（参见文献袁先智等[29]），通过马尔科夫链蒙特卡洛模拟(MCMC) 框架下的吉布斯（Gibbs）抽样方法在给定样本误差容忍（在本文样本误差的设定为 1.18%，详见 2.4 节公式 3 及其描述) 条件下将特征提取中的 NP 问题的复杂度降低为多项式复杂度筛选，从而完成刻画企业财务欺诈的特征提取工作。

3.2 公司与财务关联的坏样本(和黑样本)风险特征信息的梳理

在我国资本市场的发展进程中也有不少上市公司财务舞弊的案例，其中很多案例都曾在资本市场中 (例如银广厦、蓝田股份等)，但是这些案例相对于上市公司群体而言仍然只是九牛一毛，难以通过机器学习方法基于这些样本数据提取出有效刻画财务欺诈风险的特征因子，为解决这一难点，本文借助与财务欺诈高度关联的风险事件(Red Flag)作为判别财务欺诈样本的依据。但是，许多与欺诈事件相关的信息属于非结构化的文本处理。比如，“图 1”是 2019 年 7 月 6 日中国证监会对上市公司“康得新”做出的处罚及禁入告知的通示。



证监会对康得新等作出处罚及禁入告知

中国证监会 www.csrc.gov.cn　　时间：2019-07-05　　来源：证监会

2019年1月，康得新复合材料集团股份有限公司（以下简称康得新）因无力按期兑付15亿短期融资券本息，业绩真实性存疑，引起市场的广泛关注和高度质疑。证监会迅速反应，果断出击，决定对康得新涉嫌信息披露违法行为立案调查。

经查，康得新涉嫌在2015年至2018年期间，通过虚构销售业务等方式虚增营业收入，并通过虚构采购、生产、研发费用、产品运输费用等方式虚增营业成本、研发费用和销售费用。通过上述方式，康得新共虚增利润总额达119亿元。此外，康得新还涉嫌未在相关年度报告中披露控股股东非经营性占用资金的关联交易和为控股股东提供担保，以及未如实披露募集资金使用情况等违法行为。上述行为导致康得新披露的相关年度报告存在虚假记载和重大遗漏。

康得新所涉及的信息披露违法行为持续时间长、涉案金额巨大、手段极其恶劣、违法情节特别严重。证监会已经向涉案当事人送达行政处罚及市场禁入事先告知书，拟对康得新及主要责任人员在《证券法》规定的范围内顶格处罚并采取终身证券市场禁入措施。下一步，我们将充分听取当事人的陈述申辩意见，以事实为依据、以法律为准绳，依法进行处罚；对涉嫌犯罪的，严格按照有关规定移送司法机关追究刑事责任。

证监会重申，上市公司及大股东必须讲真话，做真账，及时讲话，牢牢守住“四条底线”，不披露虚假信息，不从事内幕交易，不操纵股票价格，不损害上市公司利益。下一步，证监会将一如既往对上市公司信息披露违法行为依法严惩，净化市场生态，努力提升上市公司质量，服务实体经济发展，为打造一个规范、透明、开放、有活力、有韧性的资本市场保驾护航。

图 1 中国证监会对康得新处罚问询数据样例

在2017年1月到2018年12月的2年时间内，有 393 家上市公司出现“黑样本”事件，这些样本基本涉及到下面用来描述财务欺诈风险场景的 11 类风险事件，这 11 种事件分类陈述如下：

（1）占用公司资产

占用公司资产指控股股东、实际控制人及其控制的其他企业利用关联交易、资产重组、垫付费用、对外投资、担保、利润分配和其他方式直接或者间接侵占上市公司资金、资产，损害公司及其他股东的合法权益。

（2）披露不实（其他）

披露不实指上市公司披露不真实的信息，即上市公司进行不真实记载、误导性陈述以及重大遗漏行为。不实信息既包括其披露的财务报告中包含的不实信息，也包括其披露的自愿性信息中的不实信息。

（3）违规担保

违规担保指挂牌公司及其控股子公司未经公司章程等规定的审议程序而实施的对外担保事项。

（4）欺诈上市

欺诈上市指公司因首次公开发行股票申请、披露文件存在虚假记载、误导性陈述或者重大遗漏，致使不符合发行条件的发行人骗取了发行核准，或者对新股发行定价产生了实质性影响，受到中国证监会行政处罚，或者因涉嫌欺诈发行罪被依法移送公安机关。

（5）擅自改变资金用途

擅自改变资金用途指上市公司违背《上市公司监管指引第 2 号——上市公司募集资金管理和使用的监管要求》规定使用资金。

（6）一般会计处理不当

一般会计处理不当指上市公司会计处理方式不符合《企业会计准则》要求。

（7）虚假记载(或误导性陈述)

虚假记载是指在信息披露的文件上做出与事实真相不符的记载，即客观上没有发生的事项被信息披露文件加以杜撰或未予剔除；误导性陈述是指信息披露文件中的某事项的记载虽为真实，但由于表示存在缺陷而易被误解，致使投资者无法获得清晰、正确的认识。

（8）推迟披露

推迟披露指上市公司的信息披露没有按照规定的时间而推迟披露的行为。

（9）虚构利润

虚构利润指通过会计舞弊的手段虚增企业税后利润的行为，通常会计舞弊行为的目的往往就是增加企业净利润，从而虚增公司的经营业绩。

（10）重大遗漏

重大遗漏是指信息披露文件未记载依法应当记载的事项，以至于影响投资者做出正确决策。

（11）虚列资产

虚列资产指通过会计舞弊的手段虚增或者虚减企业资产的行为，虚列资产的违规方式通常也会同时带来虚增利润，因此二者是可以同时发生的。

在从 2017 年 1 月到 2018 年 12 月的时间段内基于 393 家公司出现的“黑样本”中，涉及到上面陈述 11 类风险事件数量统计如表 1 所示（实际上，由于部分公司可能在 2 年间被多次问询或因多种原因被问询，因此部分风险事件发生的次数可能比“黑样本”公司的数量更多）的数据结果表明：在这两年间，涉嫌占用公司资产的事件 87 起；披露不实（其他）的事件 32 起；违规担保的事件 60 起；擅自改变资金用途的事件 17 起；序列资产的事件 4 起；一般会计处理不当的事件 30 起；虚假记载（误导性陈述）的事件 223 起；推迟披露的事件 306 起；虚构利润的事件 15 起；重大遗漏的事件 143 起。另外，在本样本中还包含 117 起未被 11 类事件包含的其他场景。

表 1 违规事件数量统计

占用公司资产	擅自改变资金用途	违规担保	欺诈上市	披露不实（其他）	序列资产
87	17	60	0	32	4
一般会计处理不当	虚假记载（误导性陈述）	推迟披露	虚构利润	重大遗漏	其他
30	233	306	15	143	117

3.3 基于人工智能算法针对刻画财务欺诈风险的特征提取和表现的上市公司实证分析

以常用财务比率，百分比报表，财务报表科目增长率为出发点构建了 183 个特征的初始特征池（见下面表 3 示例中有部分信息），利用吉布斯随机搜索方法，从中筛选出 8 个刻画财务欺诈风险的特征因子，用于建立刻画财务欺诈风险的模型，见表 2：

表 2 刻画财务欺诈的高关联特征因子

序号	特征	模型系数	p 值	关联显著性	Odds_Ratio
0	常数项	-2.49	0.00%		0.08
1	扣非净资产收益率	-0.41	0.00%	82.30%	0.67
2	在建工程增长率	-0.16	0.89%	87.40%	0.85
3	预付款项增长率	-0.15	1.05%	59.70%	0.86
4	其中：利息费用(财务费用) / 营业总收入	0.30	0.00%	97.90%	1.36
5	投资净收益 / 营业总收入	-0.15	0.24%	52.90%	0.86
6	其他收益 / 营业总收入	-0.20	0.06%	98.45%	0.82
7	其他应收款(含利息和股利) / 总资产	0.20	0.00%	99.80%	1.22
8	长期借款 / 总资产	-0.17	0.14%	65.05%	0.84

表 3 初始特征因子示例

序号	特征	关联显著性
1	实收资本(或股本) / 总资产	31.95%
2	非流动资产周转天数	25.70%
3	资本杠杆率	20.15%
4	净资产 EBIT 率	18.45%
5	短期借款 / 总资产	15.55%
…	…	…
181	应收账款周转天数	0.05%
182	其他应付款增长率	0.05%
183	吸收投资收到的现金 / 筹资活动现金流入小计	0.05%

在大数据算法过程中，本文使用的样本数据是在2018年上市的3459家公司，并按照其中80%作为训练集，20%作为测试集 (保持训练集和测试集中的黑白样本比例相同)，通过逻辑回归模型并结合比值比（Odds Ratio）对各个特征因子与财务欺诈风险的关联性强弱进行分类，得出表 2 的 8 个特征来刻画企业的财务欺诈风险。

由表 2 中可见，“扣非净资产收益率；在建工程增长率；预付款项增长率；利息费用(财务费用) / 营业总收入；投资净收益 / 营业总收入；其他收益 / 营业总收入；其他应收款(含利息和股利) / 总资产；长期借款 / 总资产”这 8 个特征指标与公司财务的其它科目有较强的关联性。

这说明在涉及到公司业务收入、税务、公司会计政策 (例如资产减值计提标准) 等因素的财务报表科目与在审计中不易查实的科目 (如预付款项、应付账款) ，财务欺诈风险具有较高的关联性。另外，静态分析和同业比较方法构造的初始特征也与财务欺诈风险具有较高的关联性，说明在机器学习框架下，静态分析和同业比较能够更有效地将上市公司的财务欺诈风险凸显出来。以8个用来刻画财务欺诈的风险特征因子为基础，图 2 中的 ROC 测试表明，基于训练集和测试集的 AUC 值分别为 0.771 和 0.766，这些数据表明我们筛选出的特征能够比较有效地刻画上市公司财务欺诈风险。

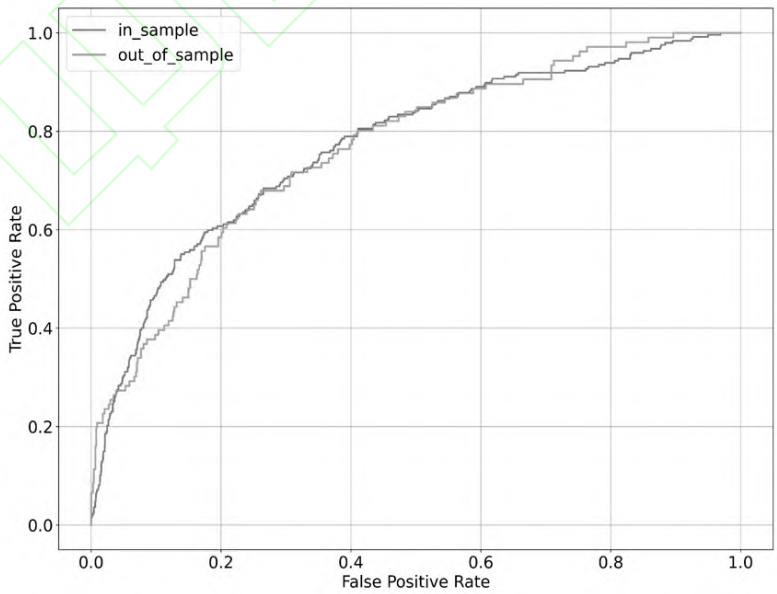


图 2 刻画财务欺诈风险模型的 ROC 测试

4 全面刻画公司财务欺诈特征的汇总与一般结论

在本节，基于舞弊审计准则（SAS99）财务欺诈舞弊的“舞弊三角理论”为出发点，结合公司董监事会的治理框架，梳理和汇总管理层是否有机会参与财务报表进行舞弊的行为表现；并结合前面讨论的刻画财务欺诈风险指标，建立有效的财务欺诈风险预警和管理。首先陈述支持刻画欺诈风险特征提取的计算表现。

4.1 支持刻画公司财务舞弊特征提取的计算表现

（1）数据与初始特征描述

1) 黑白样本数据：从 CSMAR 违规处罚数据中筛选证监会、交易所因为上市公司在 2017、2018 年因为财务报告披露不规范或真实性存疑而发出的问询函件数据，将被问询的上市公司作为黑样本。其余在 2019 年 1 月 1 日以前上市的公司中若未在以上两个年度中被问询则作为白样本。

2) 特征数据的核心指标：本文采用的特征构造方法基于财务报表的勾稽关系出发，利用公司在粉饰财务一部分科目时可能引起财务报表其他科目数据异常来进行财务报告异常的识别，因此主要特征的基础数据为“上市公司的主要财务比率、各个财务报表科目的同比增长率、百分比报表三个部分”。

（2）刻画财务欺诈风险特征提取的数值表现

特征提取的迭代效果：经过 2000 次的吉布斯随机搜索(Gibbs)迭代计算，在迭代开始时模型采用``BIC" 快速下降算法为标准，其对应的随机模拟输出结果数值一直稳定在给定的范围内（平均值为 2039.50，标准差 6.55），这说明基于 BIC 判断标准的吉布斯随机搜索(Gibbs)算法结果是稳定的。

1) 筛选特征的甄别能力结果表现：选取关联显著性指标高于 0.5 的特征作为建模特征进行建模，得到如图 3 所示的 8 个特征都与上市公司的财务舞弊风险存在显著的关联性。再结合模型的 ROC 表现（图 2）可见，模型能够有效地甄别出具有较财务舞弊风险高风险样本公司，样本内外的 AUC 值分别为 0.771 和 0.766。同时，这 8 个指标也从正负二个方面来刻画公司财务是否真实的风险：比如，“利息费用占营业总收入的比例和其他应收款占营业收入的比例和其他应收款占总资产的比例”与上市公司的财务舞弊风险存在显著的正向关联性；而其余特征如“扣费净资产收益率、在建工程增长率、预付款项增长率、投资收益和其他收益占营业总收入的比例、长期借款占总资产的比例”则于上市公司的财务舞弊风险存在显著的负相关性。

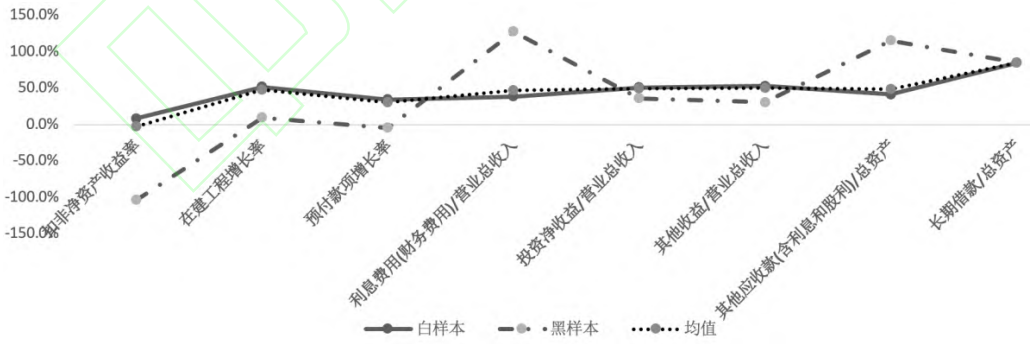


图 3 财务欺诈黑白样本平均值比较图示

2) 欺诈风险的 8 大特征对中国 A 股市场的显著性表现：基于本文筛选出用于刻画财务欺诈的 8 大特征，结合 3459 家上市公司样本 2018 年年报信息，通过针对指标的离散化分析，得到图 3 所示的高关联特征指标值（其中黑样本公司 353 家，白样本公司 3196 家）。图 3 的数据也表明，除“长期借款/总资产”对应的黑白样本差异显著性值较小外（此指标包含在 8 大指标中的一个本质原因是公司的长期债务比是公司稳定运行的一个核心基础标杆），其余 7 项筛选的特征从数值的绝对值上都能体现出黑白样本的显著差异性，这表明我们筛选出的 8 个特征指标能够对公司财务欺诈现

象的甄别,针对刻画财务欺诈风险模型的 ROC 测试也表明这些指标有比较有效的预测能力（对应的样本内和样本外的 AUC 值都在 0.76 左右），即本文筛选出的 8 个特征可以有效的支持公司财务欺诈行为的探测与预测（Detecting and Predicting）功能的落地实现。

4.2 公司监事会人数多少与公司财务欺诈的关联性关系

以 3459 家上市实体公司为样本，在公司具有比较合理的监事会人数情况下（即介于 5 到 9 人之间的情况下），测试结果表明：公司监事会人数多少与公司财务欺诈无本质关联。

统计数据分析和测试结果也表明，在公司具有比较合理的监事会人数情况下（即介于 5 到 9 人之间的情况下），公司监事会人数多少除了与公司资质无本质关联性外，对于一般的实体企业，不管是处于 A 类（从 A，AA 到 AAA 的信用评级），或 B 类（从 B，BB 到 BBB 的信用评级）或 C 类（从 C，CC 到 CCC 的信用评级）的公司，在一般情况下，25%左右的公司其董事会成员为 7 人，46%左右的公司其董事会成员为 9 人，二者相加表明，70%左右的公司其董事会成员为 7 或者 9 人。同时，对于监事会，80%左右的公司其监事会成员为 3 人，另外 14%左右的公司其监事会成员为 5 人。

实体公司（不包含金融机构）董事会平均人数与信用资质（信用评级）的关系

公司等级	平均人数	5	6	7	8	9	10	11	12	13	14	15	17	18	总计
AAA-AA	8.35	7	9	35	13	61	2	9	5	1	0	1	0	0	143
A	8.39	19	20	116	36	225	10	18	9	1	5	3	0	0	462
BBB	8.37	43	31	217	70	402	22	46	19	1	0	7	1	0	859
BB	8.52	26	27	135	68	295	13	48	11	3	3	8	0	0	637
B	8.43	32	30	172	77	359	16	43	16	2	5	4	0	0	756
CCC-C	8.23	41	36	186	55	326	11	32	10	1	1	5	0	1	705
总计		168	153	861	319	1668	74	196	70	9	14	28	1	1	3562

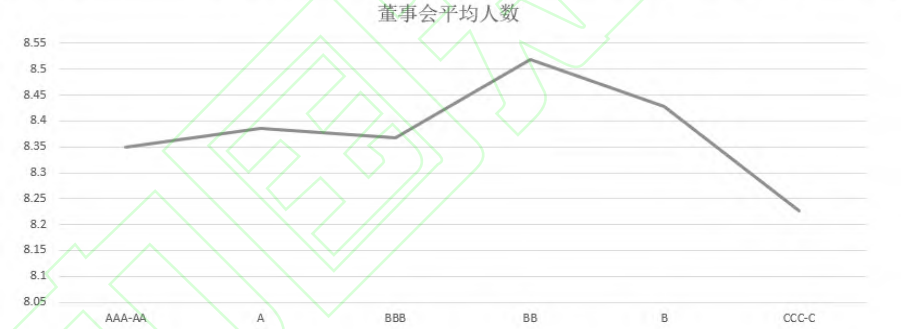


图 4 公司董监事会人数与公司资质关系

5 结语

上市公司的财务欺诈风险不仅对股东利益会造成巨大损害，也可能因为其自身的商务和规模等因素的连锁反应而引发系统风险。在业界实务实践中，公司的财务欺诈风险识别一般需要从会计，财务，法务，税务，内控管理等多个方面进行系统的分析和尽职调查，在金融科技快速发展的今天，在大数据的框架下对企业的经营，财务，金融，生态等多个维度进行“全息画像”的融合处理，在提供更加全面的尽职调查和风险评估的信息外，还提高和完善了对财务欺诈的识别与风险管理的处理能力。

本文在大数据框架下利用吉布斯随机搜索（Gibbs）方法为工具，提出了基于上市公司财务报表数据分析的财务欺诈特征提取方法，解决了由于财务报表勾稽关系而产生的维数灾难问题，从财务报表数据以及各财务数据的两两交互项中提取出 8 个特征因子来有效的刻画上市公司的财务欺诈风险，实证结果也显示本文方法能够比较有效的对刻画财务欺诈的特征进行提取，为审计，合规，投资分析等场景下的尽职调查和风险分析提供有效的技术支持。

在实际场景中，基于大数据和人工智能算法，以财务准则 No.99 号（SAS.99）标准为基本框架，将“舞弊三角理论”结合结构化和非结构化信息，

充分利用主体公司各个维度的信息，可以实现对公司财务欺诈风险的甄别与预测，并且建立动态的评估风险指标，支持动态预警与业务管理。

更进一步，需要从公司董监高的治理框架入手，结合发生财务欺诈坏样本的历史信息，和基于非结构化的 11 分类的描述，特别要思考如何充分利用深度学习，找出刻画公司在下面 3 类信息与财务欺诈的本质特征关系：1）公司审计委员会（有效性管理，如开会（解决问题）的频率）；2) 内部审计委员会成员和其有效性工作；和 3）独立的外部董事成员数和工作的有效性信息等，这是我们接下来科研工作的重点。

参考文献：

[1]BeneishMD. The Detection of Earnings Manipulation[J]. Financial Analysts Journal, 1999, 55(5):24-36.

[2] Palepu KG, Healy PM, Bernard VL, et al. Business Analysis& Valuation: Using financial statements[D]. South-Western College Publishing, 2000.

[3] AICPA. Statement of Auditing Standards No.99:Consideration of Fraud in a Financial Statement Audit[M].New York:AICPA 2002

[4] NiuG, Yu L, Fan GZ, Zhang D. Corporate fraud, risk avoidance, and housing investment in China[J]. Emerging Markets Review, 2019,(39):18-33.

[5] Healy P M, Palepu KG. Information asymmetry, corporate disclosure, and the capital markets: A review of the empirical disclosure literature[J]. Journal of Accounting and Economics. 2001,(31): 405-440.

[6] Defond M L, Zhang J. A Review of Archival Auditing Research[J]. Social ence Electronic Publishing, 2014, 58(2–3):275-326.

[7] Donovan J, Frankel R, Lee J, et al. Issues raised by studying DeFond and Zhang: What should through forensic audit researchers do?[J]. Journal of Accounting and Economics, 2014,(58): 327-338.

[8]Yang C H , Lee K C . Developing a strategy map for forensic accounting with fraud risk management: An integrated balanced scorecard-based decision model[J]. Evaluation and Program Planning, 2020, 6(80):101780.

[9] Vanhoeyveld J , Martens D , Peeters B . Value-added tax fraud detection with scalable anomaly detection techniques[J]. Applied Soft Computing, 2019, (86):105895.

[10]Nurhayati. Revealing and building the COSO concept and Khalifatullah Fill Ard philosophy to prevent and detect the occurrence of fraud through forensic accounting[J]. Procedia - Social and Behavioral Sciences 2016, (219): 541 – 547.

[11]Goode S, Lacey D. Detecting complex account fraud in the enterprise: The role of technical and non-technical controls[J]. Decision Support Systems, 2011, 50(4):702-714.

[12] Beasley M. An empirical analysis of the relation between the Board of Director composition and financial statement fraud[J]. The Accounting Review. 1996, 71(4): 443-465.

[13] 王济川, 郭志刚. Logistic 回归模型[M]. 北京：高等教育出版社, 2001.

[14] Paz A, Moran S. Non-Deterministic Polynomial Optimization Problems and Their Approximations [J]. Theoretical Computer Science, 1981,15(3): 251-277

[15]陈竞辉, 罗宾臣.亚洲财务黑洞:致命弱点在于公司治理[M]. 北京: 机械工业出版社, 2015.

[16]叶金福. 从报表看舞弊: 财务报表分析与风险识别[M]. 北京: 机械工业出版社, 2018.

[17]刘姝威. 上市公司虚假会计报表识别技术(珍藏版)[M]. 北京: 机械工业出版社, 2013.

[18] Yuan G X, Wang H. The general dynamic risk assessment for the enterprise by the hologram approach in financial technology[J]. International Journal of Financial Engineering. 2019, 6(01):1950001.

[19]王昱,杨珊珊.考虑多维效率的上市公司财务困境预警研究[J/OL].中国管理科学:1-12[2020-12-24].https://doi.org/10.16381/j.cnki.issn1003-207x.2019.1366.

[20] 洪文洲,王旭霞,冯海旗.基于 Logistic 回归模型的上市公司财务报告舞弊识别研究[J].中国管理科学,2014,22(S1):351-356.

[21] 周利国,何卓静,蒙天成.基于动态 Copula 的企业集团信用风险传染效应研究[J].中国管理科学,2019,27(02):71-82.

[22] Agrawal R., Srikant R. Fast algorithms for mining association rules[M]// Readings in database systems (3rd ed.). Morgan Kaufmann Publishers Inc. 1996.

[23]欧高炎, 朱占星, 董彬, 鄂维南. 数据科学引论[M]. 北京: 高等教育出版社, 2017, 150: 167.

[24] Qian G, Field C. Using MCMC for logistic regression model selection involving large number of candidate models[M]. In Book: Monte Carlo and Quasi-Monte Carlo Methods 2000 (edited by Fang KT, NiederreiterH, Hickernell FJ.).Springer, Berlin, Heidelberg, 2002, 460-474.

[25]Geman S. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images[J]. IEEE Trans. Pattern Anal. Mach.

Intell, 1984, 6.
[26] Qian G, Rao C. R, Sun X, et al. Boosting association rule mining in large datasets via Gibbs sampling[J]. Proceedings of the National Academy of Sciences,2016, 113(18): 4958-4963.
[27] Glasserman P. Monte Carlo methods in financial engineering[M]. Springer Science & Business Media, 2013.
[28] Narisetty N N, Juan S, Xuming H. Skinny Gibbs: A Consistent and Scalable Gibbs Sampler for Model Selection[J]. Journal of the American Statistical Association, 2018:1-40.
[29] 袁先智,刘海洋,周云鹏等.基金关联特征挖掘的大数据随机搜索算法及应用[J]. 管理科学,2020, 33(6):41-53.
[30] Akaike H. A new look at the statistical model identification[J]. IEEE Transactions on AutomaticControl, 1974,19(6):716-723.

A Feature Extraction Method on Corporate Financial Fraud

George YUAN[1,2,3],ZHOU Yun-peng[3],YAN Cheng-xing[3], LIU Hai-yang[3], QIAN Guo-qi[4],WANG Fan[2], WEI Li-jian[2],LI Zhi-yong[5],LI Bo[6], David LI[7],ZENG Tu[3]

(1. Business School，Chengdu University，Chengdu 610106，China;
2. Business School，Sun Yat-sen University，Guangzhou 510275 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. School of Finance，Southwest Univ.of Finance and Economics，Chengdu 611137 China;
6. College of Science，Chongqing University of Technology，Chongqing 400054 China;
7. Shanghai Advanced Institute of Finance，Shanghai 200030 China)

Abstract: By employing the Gibbs sampling skill under the Markov Chain Monte Carlo (MCMC), weestablish a general framework for corporate financial fraud detection by using fintech method related big data analysis. In the empirical analysis, based on those event “bad” samples from Chinese A-share listed companies enquired by China Securities Regulatory Commission (CSRC)due to behaviors such as violating (at least potentially violating) the rules of the disclosure during time period from the beginning of year 2017 to the end of year 2018 under the Rule of the Disclosure from CSRC,we conducted the analysis forkey risk factors which could represent the information for the exposure of financial fraud behavior bydetecting the difference between their financial reports from others. In general, the feature extraction (or variable selection) from around two hundred related factors of financial reports will be a NP problem because of the diversity of financial ratio indexes. However, in this paper by employing the Gibbs sampling method under MCMC, we extraction 8 key factors which are highly correlated with the behavior of corporate financial fraud, they are: ROE, the growth construction-in-process, the growth of advance payment, interest expense / revenue, investment income / revenue, other income / revenue, other receivables / total assets, andlong term loan / total assets.
The key contribution of this paper is that we establish a general framework for the extraction of key risk factors which could be used not only to detect the behavior of financial fraud, but also to predict the financial fraud under the supporting of ROC testing numerical results based on more than 3,500 A share listed companies in China.

Key words: big data; Gibbs sampling; stochastic search; SAS99; financial fraud; fraud triangle theory;theframework of feature extraction