大数据背景下税收风险识别精准度存量研究

——基于机器学习的视角

姬颜丽　　王文清

内容提要：税收风险识别对税收风险管理至关重要。无论是业务驱动的传统税收风险识别还是数据驱动的分析模型，都很难从单方面来提高税收风险识别的精准度。利用新技术提高税收风险识别的精准度成为当务之急，机器学习成为可选项。本文认为，利用风险应对核查成果的效力，把机器学习运用到税务机关防范税收风险的识别中，可实现人工经验和大数据分析的双轮驱动，为提升税收智能化管理开辟新的视野路径。本文选取随机森林算法，以商贸企业为例，对虚开增值税发票风险建立识别模型，通过检验推演验证得知，该模型稳健可靠、预测准确性高，可供税务机关参考借鉴。

关键词：大数据　机器学习　税收风险识别　随机森林　虚开发票

中图分类号：F812.45　　**文献标识码：**A　　**文章编号：**1003-2878（2020）09-0119-11

DOI:10.19477/j.cnki.11-1077/f.2020.09.010

一、引　言

自 2009 年国家税务总局提出在税收征管中引入税收风险管理理念，各地税务机关逐步设置税收风险防控机构，建立风险应对的识别机制，确立风险管理的操作流程，取得一定成效。税收风险识别模型是税收风险管理的重要工具，支撑着风险识别、风险排序、风险应对等风险管理流程的关键环节。因此，税收风险识别模型的科学性和准确性决定税收征管资源能否得到合理有效配置，影响着税收风险管理的成效与结果。然而，随着市场主体急剧增加，新经济、新业态、新商业模式不断涌现以及税收政策的不断变化等，都给税收风险管理带来新挑战。面对新形势，国家税务总局提出要充分利用税收大数据，并大力开展“智税”竞赛，推动各地税务机关积极探索更加科学、精准的税收风险识别模型。为此，研究

作者简介：姬颜丽，国家税务总局税务干部学院，讲师。
王文清（通讯作者），国家税务总局税务干部学院，教授。

基金项目：本文系国家税务总局税务干部学院青年课题“大数据技术背景下完善税收风险管理体系建设的研究”（项目编号：2020QNKT005）的阶段性研究成果。
作者感谢匿名审稿专家所提宝贵建议，当然文责自负。

如何利用税收大数据提高税收风险识别模型的精准度至关重要。本文对税收风险识别精准度存量研究专指税务机关对纳税人在纳税申报、税收管理以及风险防范等方面可以及时预测、识别、发现问题的精确度与可靠性，并能够从更深、更广、更贴近实际的角度去开展符合业务需求的存量思考与研究。

目前，国内对于大数据在税收风险管理中的作用探讨较多。姚键等（2015）立足税务实践，指出目前税收风险分析识别停留在简单查询和比对层面，缺乏有效的数据分析工具。税务机关应强化数据分析和数据挖掘，以大数据技术推动智能化管理。常晓素（2019）、宋星仪和宋永生（2020）认为，税务系统未能充分利用其拥有的海量涉税信息，税收风险识别命中率不高，税收风险管理效率低下，应培养税务人员的大数据思维，强化“以数治税”，创建全方位税收风险管理体系。学界普遍认可大数据对税收风险管理的重要作用，以及数据分析对于税收风险管理的重要意义，但是专门研究如何通过数据分析提高税收风险管理效能的较少。理论层面的探讨有：张景华（2014）以行业税收风险识别模型为例，提出从投入产出类指标和财务税收类指标构建模型，通过预警值和权重确定风险得分。李为人和李斌（2018）、刘昊（2020）提出在税收风险分析中引入机器学习智能算法，可自主建立指标模型，并通过不断试错，形成行之有效的模型，从全景分析中描绘风险分布情况、评估风险轻重缓急，可以实现客观、精准的分级分类管理，不断提高纳税遵从。技术层面的探索有：刘尚希和孙静（2016）探索运用大数据进行税收风险识别，并选取机器学习中的神经网络模型，来识别某市软件行业企业所得税风险。文章受数据所限，仅选取 7 家有问题企业，无法涵盖风险企业的全部特征。但作者为以后建模提出了改进的方向，扩大样本数量和特征值范围，并比较机器学习中不同算法的适用性来选取模型。夏会等（2019）选取 63 家企业的财务和纳税数据，使用无监督的机器学习模型——K 均值（K-means）聚类算法，对房地产企业股权转让的税收风险进行聚类分析。但聚类的结果不能表明哪类是高风险企业，需要税务人员结合经验判断得出。基于先前研究，在大数据背景下将机器学习应用于税收风险识别成为必然趋势。然而，由于缺乏风险应对结果数据，无法比较机器学习中不同算法在具体风险识别问题中的适用性，也无法应用到风险识别模型的迭代中。有鉴于此，本文在剖析当前税收风险识别模型和机器学习构建识别模型各自不足和优势的基础上，提出充分利用风险应对核查结果，以商贸企业增值税虚开发票风险识别为例，借助有监督的机器学习构建风险识别模型，实现人工经验和大数据分析的双轮驱动，提高税收风险管理的效率和效能，推动“互联网 +”在税务领域的深度融合。

二、税收风险识别模型的现状及存在的问题

2018 年国地税机构合并之后，各省新成立的大数据和风险管理机构，依托金税三期税收大数据管理平台，全面归集各类数据，整合原有风险指标和风险模型，建立了覆盖全税种、全行业以及特定事项的税收风险识别指标和模型，实行全省统一，分级分类风险应对管理。税收风险识别模型是发挥税收大数据价值的关键，也是关系税收风险管理效能的核心。但是，当前税收风险识别模型主要基于小样本的典型调查设计风险指标，再结合预警值对风险进行识别，后运用综合评价法设置权重，计算出综合风险得分，进行风险排序。其公式为：风险得分 $=\sum$ 指数权重 × 单项风险得分。这种单一、传统的业务判定模式，在税收风险识别方面过于依赖建模者的经验水平，无法充分利用税收大数据的优势，识别精准度不高，难以满足大数据背景下的税收风险管理需要。

（一）税收风险识别指标的针对性不强，风险识别度低

识别指标是税收风险识别模型的关键要素。各地税务机关虽建立了覆盖全税种、全行业的税收风险识别指标。但指标库中多是税负率、税收增长率、财务指标等较为简单的通用指标，针对性不强，

缺乏特色和复杂度。一方面，当前指标设计主要依靠建模者的经验研判，通过对申报表和财务报表的简单比对和分析得出，存在局限性。对于新产业、新业态、新商业模式，建模者难以了解其风险特征，设计的指标无法触及风险实质。对于一些经营模式复杂的行业，建模者缺乏深入调研分析，未能触及行业深层问题，设计的指标无法识别出重大风险。对于虚开发票、股权转让等复杂事项，目前缺乏有效的风险识别指标，利用税收大数据资源的韧度不足，无法精准定位税收风险。如对 Y 市正常领用、开具“通用机打发票”的 7960 户纳税人抽样数据结果显示，“纳税人使用通用机打发票”等 6 类疑点问题中“发票作废异常”和“负数发票异常”两个疑点均在实地核查后被排除，由于人力、物力及时间所限，精准识别度达不到 75%。可以看到，指标的数据主要取自纳税人税收和财务信息，未能充分利用与纳税人相关联的其他信息，发挥数据价值。因此，当前税收风险识别指标体系还不够完备，需要进一步扩展指标的数据来源，从更多渠道选取指标，来增强指标体系的针对性和有效性。

（二）税收风险指标的参数设置不够科学，风险排序合理性差

预警值和权重是税收风险识别模型的重要参数，影响风险识别效果和风险排序结果。建模者在实践中通常根据小样本的典型调查主观确定预警值，与实际存在较大偏差，使得识别出的纳税人有不存在相应风险和有相应风险却未被识别出来的情况。同时，同一行业内不同企业性质、规模或者核算模式在风险指标取值上可能存在较大差异，仅通过简单平均得到的居中预警值，无法应对差异、难以筛选到真正的风险企业。权重的确定是当前税收风险识别模型的难点，需要利用统计分析方法，结合定量和定性来确定。建模者在确定权重时常因无取数权限或者缺乏数据分析技能而自主赋权，随意性强。目前，模型参数的设置还没有明确的量化标准和确定方法，更无绩效考核要求。税务机关在预警值和权重设置上，主要以纳税人涉税排查的实践经验为主，缺乏科学、合理、规范的检验标准。如对纳税人由于税负率偏低或财务指标异常来判定税收风险的存在，并要求纳税人进行税收调整等。

（三）模型修正成本过高，与风险应对需求结合不紧密

当前税收风险识别模型过度依赖建模者主观经验，无法形成客观化、规范化的标准，导致模型推广和复制的局限性很大。风险应对通常以查补税款为评价标准，缺乏提供模型修正建议的有效激励措施。加之，传统税收风险识别模型无法吸收微观风险应对案例数据，只能通过分析总结后的要点反馈调整预警值或权重，进而以此修正模型。如果风险应对结果普遍显示模型效果不佳，则需要重新设立指标，建立模型，无法迭代调整，修正成本极高。税收风险识别模型和税收风险主要为单向纵联关系，二者互动性小，结合不够紧密。如，税务机关风险应对人员所提供的风险数据，对分析人员所起到的实用价值就微乎其微。也正如此，风险识别模型一经设立，在后续的使用与更新中，缺乏有效的反馈与修正机制，难以应对不断变化的税收风险。

三、运用机器学习建立税收风险识别模型的机遇和优势

随着国地税征管改革的深入，金税三期系统在原国税、地税机构合并后实现流程统一，数据合并和功能升级。税务机关拥有包括核心征管、个人税收管理、电子底账以及出口退税等涉税数据以及第三方数据，掌握从纳税人登记到注销整个活动的数据信息。税务机关集聚海量的涉税数据，亟需通过更深、更广的数据分析技术来发挥数据价值。运用机器学习技术进行税收风险识别顺应了大数据背景下税收风险管理的迫切需求。机器学习以计算机可利用的数据信息为驱动，利用计算机科学的基础概念，结合统计、概率和最优化的思想来得到模型，做出更准确的预测。它致力于从经验数据中学习算法来改善系统性能，用于对新情况做出有效决策。风险应对成果包含较为全面的企业风险信息和风险特征，是风险应对人员

业务分析和调查取证的结晶。当前，税务机关逐步从风险疑点消除式应对转变为全面、专业的团队式应对，这为机器学习积累了丰富的标识数据，可用来训练模型、提高模型精准度。机器学习是人工智能的技术基础，把机器学习运用到税务机关防范税收风险的识别中，可开辟提升税收智能化管理的新路径。

（一）利用大数据思维从相关视角选取指标，识别更有效

机器学习利用大数据思维从相关关系视角来选取指标，从传统的因果分析扩展到对纳税人行为模式的分析，极大扩展指标选取的范围，能更有效利用税务机关掌握的涉税数据。机器学习满足大数据分析需求，支持全样本、全变量参与，根据相关性自动选择特征变量，识别更精准，效率和准确性都高于人工。如此，机器学习就易于从大量数据中发现规律，对于虚开骗税、股权转让等复杂事项的分析和处理，具有人工经验无法比拟的优势。特别是，恶意的涉税风险通常经由不法分子精心设计，隐蔽性强，需要大数据和多指标才能找到蛛丝马迹，而这恰是机器学习的长处。对于机器学习来说，训练样本的数量越大，特征变量维度越高，算法学习效果越好，识别越有效。

（二）利用丰富的算法自动建立识别模型，识别更智能

机器学习集成类型丰富、数量众多的算法，如决策树、神经网络、支持向量机、随机森林、逻辑回归等，能用于数据分析和预测。在构建税收风险识别模型时，机器学习能根据企业不同的特点和规律，从企业规模、企业类型、经营方式、管理模式、核算方式、技术水平等多维度来设立算法，自动训练函数，建立识别模型，突出企业的风险特征。传统税收风险识别模型属于综合评价问题，需要人工确定每个指标的预警值及其权重，成本高、效率低，质量标准难统一。然而，机器学习可针对具体的风险问题，使用“最优化”的思想，选择最佳的算法，自动调参，建立模型，更有效率，更加智能。

（三）利用监督学习算法使模型迭代升级，识别更精准

机器学习分为监督学习、无监督学习和强化学习。通过监督学习算法，能够充分利用风险应对案例的微观数据，有效学习人工核查经验，完善风险识别模型。运用监督学习算法可建立识别模型与风险应对的良性互动（见图 1），形成“学习—应对—再学习—再应对”的循环，不断提高识别模型的准确性和应对的高质量。风险应对结果数据，由税务人员对企业进行深度分析和检查形成的，包含较为详细的涉税风险点，可充分利用该数据来提取纳税人的风险特征，为纳税人“画像”。将税收风险应对案例数据应用于机器学习税收风险识别中，将人工经验数据化、模型化，无疑会提升税务人员核查的效率。也就是说，通过机器学习技术训练算法，能够对未核查纳税人进行精准预测，相当于将核查的经验推而广之，大大降低征管成本，提高纳税遵从。

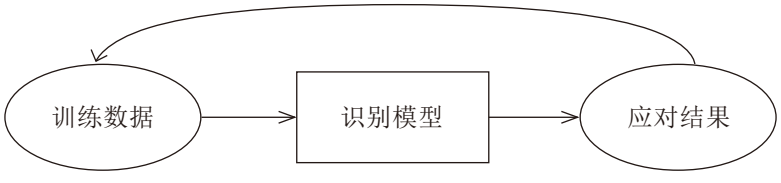

图 1　机器学习识别模型与风险应对反馈的关系

四、运用机器学习构建税收风险识别模型的实证分析

（一）问题定位

目前虚开增值税发票现象屡禁不止，严重扰乱国家的税收秩序，属于重点税收风险管理领域。其中，

商贸企业凭借其涉及的产业链条长、开办成本小、经营灵活、范围广泛等先天条件，成为不法分子作案的有力工具。从以往案件来看，往往是不法企业协同寻求虚进发票来源，再联系受票企业，通过洗票收取手续费获得虚开发票牟利，形成较为隐蔽虚受虚开发票的网络链。由于商贸企业作案时间短，不易被税务机关及时发觉，且没有固定资产的累赘，导致不法企业容易再次新办商贸企业作案。传统的税收风险识别模型主要用于事后防控。面对数量众多的商贸企业，税务机关缺乏足够的人力和资源来识别和应对。因此借助机器学习，学习以往专家风险应对成果，自主构建高精准的税收风险识别模型，及时监控商贸企业的风险状态，对于防范虚开意义重大。

（二）实证分析

本文抽取某地 1048 户商贸企业的样本，主要分为无虚开增值税发票和有虚开增值税发票两类。其中，税务核查结果为 190 家无虚开，858 家有虚开，样本足够大，且有虚开和无虚开的样本都比较多，包含各种虚开发票的特征。通过分析，发现 90% 以上的商贸企业存在无财务报表或申报表的情况。因此，通过财务报表和纳税申报表数据的因果关系进行风险识别难以进行，必须考虑利用其它相关数据信息来识别风险。识别是否虚开属于分类问题。机器学习中可解决分类问题的算法有逻辑回归、神经网络、支持向量机、朴素贝叶斯、决策树、随机森林等。对于复杂的分类问题来说，训练一个复杂的分类模型通常比较耗费时间，同时，为了能够提高对分类问题的预测准确性，通常可以选择训练多个分类模型，并将各自的预测结果组合起来，得到最终的预测。因而，集成学习对于虚开发票的分类问题更适用。随机森林是一种重要的基于装袋（bootstrap aggregating, 简称 Bagging）的集成学习算法，由 Breiman（2001）提出。本文选择随机森林算法来建立模型。

随机森林由一系列的决策树集合构成，所以称为“森林”。方匡南等（2011）指出，随机森林通过自主采样法（Bootstrap Sampling）从原始数据集中有放回地抽取多个样本，然后对每个样本分别构建决策树，结合每棵决策树的决策结果，通过投票方法得到最优结果（流程见图 2）。随机森林具备以下特点：

（1）通过自主采样法抽取子训练样本，构造不同的训练集增加了分类模型间的差异。假设训练样本集用 S 表示，测试样本集用 D 表示。每次抽取的训练样本集为 S_k（k=1,2,…, K），各样本集相互独立，具有相同的分布。S_k 的样本量与 S 的样本个数相同，假设都为 N。S 中，每个元组被抽中的概率为$\frac{1}{N}$，抽取 N 次后某个元组未被抽中的概率为 $(1-\frac{1}{N})^N$, 当 $N\to\infty$ 时，$(1-\frac{1}{N})^N\to\frac{1}{e}\cong 0.368$。在每轮随机抽样中，训练样本中大约有 36.8% 的样本没有被抽到过，这部分数据称为袋外数据（Out Of Bag，简称 OOB）。袋外数据可以检测模型的泛化能力。因此，随机森林泛化能力很强，可以有效降低模型的方差。

（2）随机森林泛化误差具有收敛性。假设通过训练集 S_k 得到决策树的分类模型为 $h_k(\mathbf{X})$，定义其余量函数（Margin Function）来度量平均正确分类数目超过平均错误分类数目的程度：

$$mg(\mathbf{X},Y)=av_k I(h_k(\mathbf{X})=Y)-\max_{j\neq Y} av_k I(h_k(\mathbf{X})=j)$$

余量值越大，分类预测就越可靠。泛化误差可写成：

$$PE^*=P_{\mathbf{X},Y}(mg(\mathbf{X},Y)<0)$$

随着决策树 K 的增加，泛化误差总收敛于$P_{\mathbf{X},Y}(P_{\Theta}(h(\mathbf{X},\Theta)=Y)-\max_{j\neq Y} P_{\Theta}(h(\mathbf{X},\Theta)=j)<0)$。

（3）每次随机选择少量的特征用来训练决策树，进一步增强了模型的泛化能力。对于全部 M 个特征，一般随机选择 $m=\log M$ 个特征来训练决策树，$m\ll M$。

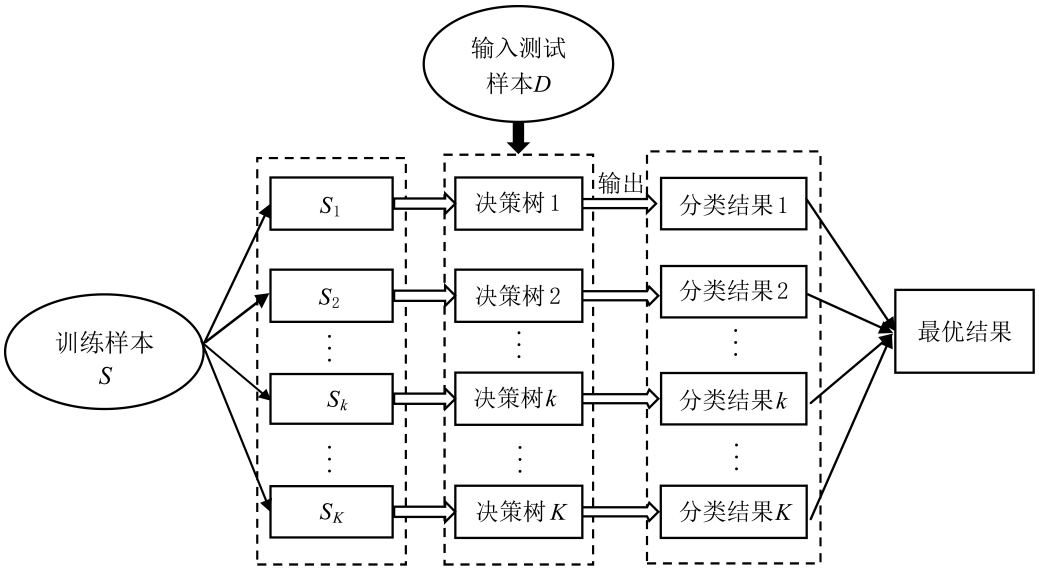

图 2　随机森林流程图

随机森林预测精度高于传统的分类模型，泛化误差更小，学习速度更快，抗噪声效果强，能够处理更高维度的数据。相比其他分类算法，随机森林优势明显，是处理分类问题优选算法之一。此外，随机森林对部分特征的缺失不敏感，对于存在数据缺失的商贸企业仍能进行有效识别。

结合商贸企业风险特征，本文从企业税务登记、税收行为、税收经济和税收风险四个方面 25 项指标观察风险企业的涉税行为特征（见表 1）。其中，X1–X4 属于税务登记类；X5–X12 属于税收行为类；X13–X17 属于税收经济类；X18–X25 属于税收风险类。指标数据来源于金税三期系统、防伪税控系统使用的风险识别指标。数据使用前进行了脱敏处理。

表 1　指标编号及名称

编号	指标名称	编号	指标名称
X1	企业资金规模	X14	月均销售收入
X2	企业人员规模	X15	增值税税负
X3	企业法人户籍	X16	月均销售收入相对行业平均差
X4	企业成立时间长短	X17	增值税税负相对行业平均差
X5	发票领用量月离散度	X18	票表信息不符
X6	发票平均开票金额	X19	购销货物品名不匹配
X7	发票销售对象数量	X20	有销售无购进
X8	发票销售对象区域	X21	大量抵扣低税率专用发票
X9	发票受票对象数量	X22	从同一家上游公司购进货物
X10	发票受票对象区域	X23	大量抵扣农产品发票
X11	发票开票商品种类数量	X24	自开增值税普通发票数额较大企业
X12	月均领票频度	X25	同址同经营范围
X13	企业首次实现税收时间		

首先，将 1048 户企业样本随机分为两部分：一部分是训练数据，用来训练识别虚开发票的模型；另一部分是测试数据，用来评价模型的泛化能力。本文使用基于 Python 语言的机器学习工具 Sklearn 库 (Scikit-learn) 来进行建模计算。首先将原始数据随机划分为训练集和测试集，一般选取 75% 的数据为训练集，25% 的数据为测试集。然后对训练集构建随机森林分类树模型。随机森林分类树模型必须确定以下关键参数：（1）森林中决策树的个数 K。一般来说 K 太小，容易欠拟合，K 太大，计算量太大，且 K 大到一定程度对于模型的提升会很小，所以需要选择一个适中的数值。（2）决策树划分时对特征的评价标准。评价标准有基尼指数 (Gini index) 或者信息增益。（3）单棵决策树分割节点时使用特征的最大数量。这里选择总特征的平方根，即 5 个。（4）决策树最大深度。如果样本量少或者特征少，则不限制最大深度。如果模型样本量多，特征也多的情况下，推荐限制最大深度。对于参数的确定，采用网格搜索（Grid Search）的方式来进行调参。利用网格搜索，在所有可能的参数范围中，通过循环遍历，尝试每一种可能性，表现最好的参数选为最终的结果。网格搜索会计算出模型多个参数的所有可能组合，并列出最好的参数。通过网格搜索，得到最优参数为：决策树的个数为 22，节点特征选择方法为基尼指数，决策树最大深度为 8 。最优参数得分为 0.8511，该参数组合表现最佳。

输入训练样本，利用上文得到的最优参数，构建随机森林模型。随机森林特征重要性表现为在该特征上拆分的所有节点上基尼指数的平均减少。将基尼指数的平均减少值进行降序排列，可以得到指标的重要性排序（见表 2）。重要性排名前五的指标依次为 X13、X7、X17、X11、X4。重要性排名后五的指标依次为 X21、X18、X19、X23、X22。

表 2　指标重要性度量

编号	基尼指数的平均减少	重要性排序	编号	基尼指数的平均减少	重要性排序
X13	0.1004	1	X5	0.0342	14
X7	0.0916	2	X6	0.0329	15
X17	0.0849	3	X25	0.0321	16
X11	0.073	4	X12	0.0282	17
X4	0.0713	5	X3	0.0207	18
X15	0.0707	6	X20	0.0086	19
X14	0.0609	7	X24	0.0075	20
X16	0.0505	8	X22	0.0062	21
X2	0.045	9	X23	0.0057	22
X9	0.0436	10	X19	0.0055	23
X10	0.0425	11	X18	0.0045	24
X8	0.0406	12	X21	0.0004	25
X1	0.0387	13			

通过指标的重要性排序，可以发现排名靠前的指标以税收经济类为主，其次是税收行为类和税务登记类。税收风险类指标除了 X25 是否同址经营外，其他的指标 X18-X24 分布于排名最后 7 位。

根据优化后的模型，拟合测试样本，得到预测结果。对于二分类问题，可将样例根据其真实类别与模型预测类别的组合划分为真正例（TP, true positive）、假正例（FP, false positive）、真反例（TN, true

negative）、假反例（FN, false negative）四种情况，令TP，FP，TN，FN分别表示其对应的样例数，则可以得到TP+FP+TN+FN=样例总数。分类结果的混肴矩阵（confusion matrix）如表3所示。

表3　混淆矩阵

实际	预测	
	正例	反例
正例	真正例（TP）	假反例（FN）
反例	假正例（FP）	真反例（TN）

评价预测的准确性有许多指标，其中最常见指标有：准确率（Accuracy）、精确率（Precision）、召回率（Recall）、F测度值（F-measure）和AUC(Area Under ROC Curve)。其中准确率为正确分类的样本数占总样本数的比例；精确率是分为正例的样本中真正为正例的比重；召回率是正例样本被分为正例的比重；F测度值为精确率和召回率的调和均值。AUC表示受试者工作特征 (Receiver Operating Characteristic, ROC) 曲线下的面积，介于0.1和1之间。AUC的值可以直观的评价分类器的好坏，值越大越好。各指标的具体计算如下：

$$\text{Accuracy}=\frac{TP+TN}{TP+TN+FP+FN}$$

$$\text{Precision}=\frac{TP}{TP+FP}$$

$$\text{Recall}=\frac{TP}{TP+FN}$$

$$\text{F-measure}=\frac{2*\text{Precision}*\text{Recall}}{\text{Precision}+\text{Recall}}$$

为了检测随机森林算法对识别商贸企业虚开增值税发票预测的准确性，我们将上述介绍的指标都进行了计算，准确率为0.9513，精确率为0.9633，召回率为0.9778，F测度值为0.9705，说明模型预测准确性非常高。AUC的值为0.9047，说明该分类器分类效果很好。为了比较其他分类算法与随机森林算法效果的有效性差异，我们利用逻辑回归算法、支持向量机算法作为对比，分别计算了各个模型的评价指标。结果见表4。从准确率和精确率来看，逻辑回归算法和支持向量机算法都低于0.9，远低于随机森林算法。召回率一般与准确率呈反向变动，逻辑回归算法和支持向量机算法的召回率略高于随机森林算法。随机森林算法的F测度值远高于逻辑回归和支持向量机。而且从AUC的值来看，逻辑回归为0.6339，支持向量机为0.5638，随机森林为0.9047。除了召回率，其他评价指标都是随机森林算法最优。尤其是从AUC的结果来看，随机森林算法远远高于其他两种算法。因此，随机森林算法最优，稳健性最好，可利用该模型对商贸企业虚开增值税发票的行为进行识别。

表4　各个算法评价指标对比

评价指标	逻辑回归算法	支持向量机算法	随机森林算法
准确率	0.8273	0.8042	0.9513
精确率	0.8289	0.8187	0.9633
召回率	0.9941	0.9985	0.9778
F测度值	0.9040	0.8997	0.9705
AUC	0.6339	0.5638	0.9047

（三）结论认定

上述分析表明，在商贸企业虚开增值税发票的风险识别中，随机森林模型对于变量繁多、类型复杂、涉及业务广泛的情况具有最高的识别准确性。该模型可以广泛应用于成千上万的商贸企业增值税发票风险的识别。可见，随机森林算法在识别商贸企业增值税发票风险模型中，高效使用了税务机关核查的成果数据，将人工经验模型化，自动识别更多纳税人。由于受数据所限，标记数据仅对风险结果分了两类，未来可考虑将风险应对结果充分细化，以便能对企业进行精准画像和风险识别。税务机关可根据风险应对的结果对纳税人进行有效标记，获取丰富的样本，利用大数据和机器学习的优势，训练识别模型，对纳税人进行更精准的预测和分析能够给予税收风险管理最大助力。

通过对商贸企业虚开增值税发票的分析，可以证实机器学习模型能够有效解决传统税收风险识别的不足，即解决了因申报数据缺失造成的无法识别和识别不准确的问题。通过机器学习建立的识别模型，经测试数据的验证，准确性很高。因而可将机器学习广泛推广到其他行业的税收风险识别模型构建中。在运用机器学习构建识别模型时需要具备以下条件：一是数据质量。高质量的大数据是机器学习的前提。税务机关需一方面充分收集和存储已有的征管数据，另一方面扩展数据的来源，多渠道获取相关数据。二是算法方式。合理、准确的算法方式是解决问题的关键。税务机关可依托税收风险管理平台，部署丰富的算法库，针对建模目标广泛选取数据指标，训练合适的算法，构建应用得体的识别模型。三是计算能力。计算能力是实现税收风险防控最终目标的保障。税收风险管理平台需具备强大的计算能力，配备相应的计算芯片和服务器。同时要注重培养兼具业务能力和数据分析能力的专业化人才。

五、机器学习应用于税收风险识别的展望及建议

随着大数据在各个领域的应用与发展，它的科学性与实用性越显重要。税务机关应当把握机会，不断加强税收信息化的应用管理，采取先进的技术方法、探索未来发展方向，才能将税收风险的系数降到最低。

（一）积累应用数据信息，建立税收风险识别数据库

商贸企业虚开增值税发票案例分析说明，税务机关利用机器学习建立识别模型，有效地识别了虚开增值税发票的风险。税务机关在建立税收风险识别数据库时，一是扩展思维的宽度，从因果分析的角度扩展到大数据相关性思维的宽度，广泛利用一切可以利用的数据，建立税收风险识别数据库，尽可能采集较多的带有纳税人特征的信息。通过大量有价值的特征信息，构建算法模型，对纳税人的涉税风险进行高度准确的识别，从而使得税务机关可以集中有效资源应对风险。二是拓宽数据来源的渠道。数据是税收风险管理最重要的资源，税务机关获取的数据资源越多，对纳税人的服务就能够越精准，管理越有效。应高度重视数据管理，提高数据管理的地位，发挥数据的资源优势，全方位搜集从纳税人特征到行业方方面面的信息，抓取可以使用的数据，进行分析，发现规律。此外，对于不同的纳税人，可采取差别化数据管理。比如，对于无能力报送或者无法报送准确的标准财务报表和申报表的纳税人，可考虑利用其他数据对其进行管理，简化其数据报送内容，降低纳税人的负担。

（二）强化风险应对管理，建立税收风险防范的运行机制

在当前大数据背景下，税务机关应利用大数据技术推进税源专业化管理向高、深、尖方向发展，

并根据内部职能部门的功能重新划分岗位职责。具体说，一是加快系统整合，拓展系统功能，打通内部各系统间数据壁垒，尽快实现金税三期核心征管系统、增值税发票管理新系统和第三方交换系统的数据共享融合，形成面向各级税务机关的综合数据分析应用平台。二是通过大数据分析技术进一步细化税收预警风控系统，对重点和疑点数据全天候、全方位的实时监控，建立具有特征标识的税收风险数据库。三是加强大数据跟踪工具对税收风险特征库指标运用效应的评估，并不断根据实际指标的变化更新特征库，进行实时维护以提高应对各类税收风险的识别功能。四是强化信息统筹，构建内部联动防范机制，通过信息化手段提升税源管理、发票监控、风险分析等各环节、各层级联动防范能力。

（三）把握风险识别方法，探索税收征管数据融合的绿色通道

当前，我国税收风险管理还处在边应用边摸索阶段，并没有完全形成科学的、集约的、智能的税收风险分析识别路径，缺少纵横交错、配套关联的大数据平台技术支撑和监管一体的基础设施。因此，首先，我国各级税务机关应加强风险管理团队的建设，不拘一格的聚拢和招聘优秀人才，设立相应的税收风险识别与监管专业化团队，把大数据技术应用与税收征管有机结合，研究和探索适合我国国情和税收特色的风险防控模式。在现有条件下不断拓展新方法、新手段、新技术，进而实现税收风险管理向着高质量、高水平、高效率的方向发展。其次，建立全国统一的防控税收风险快速响应机制。通过税务总局及省局两级机关的各类发票信息比对分析，直接向基层推送税收风险的线索，减少中间层级，实现基层迅速核查。进一步优化税务系统内部不同地域间的税收风险管理联动响应机制，实现联防联控，快速传递，合力防范。再次，进一步放宽税务总局大数据云平台使用权限，以便于各级税务机关能够充分深入运用大数据云平台及时开展风险分析，为涉税违法案件的快速反应查处，提供大数据分析应用支持。

（四）依托大数据应用平台，建立税收共治的良好环境

机器学习应用于税收征管及风险防范的合理性、科学性及先进性，离不开社会对税收环境的实践与体验。因此，一是应该强化信息共享共用，构建综合治税协作机制。依托“互联网 + 政务”建设，尽快构建政府各职能部门参与的综合治税信息协作共享平台，实现互通互联、数据共享，方便税务机关应用大数据管理理念，及时有效分析锁定税收风险，最大限度发挥多部门治理涉税违法的工作合力。二是强化信用分类管理，构建纳税信用联合惩戒机制。将涉税不法企业纳入重大失信行为，建立面向社会的纳税信用应用分类机制，实现多部门联合惩戒。三是强化服务措施分类，实施风险纳税人“黑名单”管理。对于曾在非正常户企业、违法经营企业任职的企业经营人员列入风险纳税人“黑名单”，在领购发票、票面增量升版、商事企业登记注册、银行贷款等方面严格限制，重点监控，营造快速治理的环境，有效震慑的社会效果。

参考文献

[1] 常晓素 . 大数据在税收风险管理中的应用探析 [J]. 税务研究, 2019 (06) : 78–81.

[2] 方匡南 , 吴见彬 , 朱建平 , 谢邦昌 . 随机森林方法研究综述 [J]. 统计与信息论坛, 2011 (03) : 32–38.

[3] 李为人 , 李斌 . 在税收风险分析中引入人工智能技术的思考 [J]. 税务研究, 2018 (06) : 29–34.

[4] 刘昊 . 人工智能在税收风险管理中的应用探析 [J]. 税务研究, 2020 (05) : 79–82.

[5] 刘尚希, 孙静 . 大数据思维 : 在税收风险管理中的应用 [J]. 经济研究参考, 2016 (09) : 19–26.

[6] 邱吉福, 张仪华 . 基于税收执法视角下企业税收风险管理研究——以厦门市为例 [J]. 财政研究, 2012 (09).
[7] 宋星仪 , 宋永生 . 大数据在税收风险管理中的应用探析大数据环境下税收风险管理的路径选择 [J]. 税务研究, 2020 (03): 99-103.
[8] 夏会 , 程平 , 张砾 . 大数据下基于改进 K-means 聚类算法的税收风险识别 [J]. 财会月刊, 2019 (11): 143-146.
[9] 姚键 , 王周飞 , 陈爱明 . 基于大数据背景的税收风险管理 [J]. 税务研究, 2015 (11): 64-66.
[10] 张景华 . 税收风险识别模型的构建 [J]. 税务与经济, 2014 (01): 96-99.
[11] Breiman,L.. Random Forests. Machine Learning,Vol.45, No.1, 2001:pp.5-32.

The Stock of Research on Accurate Identification of Tax Risk under the Background of Big Data Technology ——Based on Machine Learning

Ji Yanli Wang Wenqing

Abstract: Tax risk identification is important for tax risk management. Neither the traditional business-driven tax risk identification nor the data-driven analysis model can improve the accuracy of tax risk identification unilaterally. Therefore, we put forward that we must make full use of the results of risk response verification, apply machine learning to the identification of tax authorities to guard against tax risks, realize the two-wheel drive of manual experience and big data analysis, and open up a new visual path for improving the intelligent management of tax revenue. In this paper, the random forest algorithm is selected to establish the identification model for the risk of false invoicing of business enterprises. Through the test, the model is robust and reliable, and the prediction accuracy is high, which can be widely used.

Keywords: Big Data Technology; Machine Learning; Tax Risk Identification; Random Forest Algorithm; False Invoicing

(责任编辑 : 高小萍)