DOI:10.19641/j.cnki.42-1290/f.2019.21.023

大数据下基于改进K-means聚类算法的税收风险识别

夏 会[1](博士)，程 平[1](博士生导师)，张 砾[2]

【摘要】目前的税收风险管控模型通常是基于税务人员的先验知识构建的，在海量数据环境下模型的实用性、可扩展性和精确性都有较大的局限。为解决这一问题，提出改进的K-means聚类算法。该方法是无监督学习模型，可以在无先验知识的前提下构建指标，快速且精确地对实例进行聚类，将出现明显异常的小类识别为异常，判定其存在税收风险。基于该模型对房地产类企业股权转让中的税收风险进行分析和评估，发现税收风险等级高的企业及其风险疑点，验证了本方法的有效性。

【关键词】税收风险；K-means聚类算法；大数据；股权转让

【中图分类号】F812.42　　【文献标识码】A　　【文章编号】1004-0994(2019)21-0143-4

随着“互联网+税务”的日益深入，以电子税务局为媒介，各省市税务机关收集了大量的纳税人相关数据。金税三期平台的成功上线和不断完善，进一步统筹了税务部门和相关涉税机构，使涉税数据呈现指数级的增长[1]。面对海量的涉税数据，依赖于先验知识的税收风险管控工作已经无法发挥优势，需要基于机器学习、数据挖掘等智能化模型对数据进行科学化、精细化分析，以及时发现纳税疑点，辅助税收征管工作。因此，研究如何应用未标注数据集构建高精度、可扩展、实用的税收风险识别模型，发现纳税疑点，实现对税收风险的有效管控具有重要的现实意义和价值。

一、文献综述

当前基于大数据对税收风险的研究主要集中在构建税收风险管理相关平台的系统和模型上。徐壁[2]从数据的角度出发，基于大数据技术，构建了税收风险管理系统，主要包括涉税大数据的采集和存储及相应的标准体系，涉税风险防控体系及相应的指标模型，以及涉税大数据分析与挖掘平台。刘小瑜等[3]则提出了针对高新技术企业的税收风险预警模型构想，并在模型中引入了智能优化算法以增强税收风险识别的精度。但由于缺乏足够的已标注的数据，该模型的实施效果有待进一步的验证。刘尚希等[4]基于某区2012年和2013年纳税申报数据和财务报表数据，提取指标，构建人工神经网络模型对纳税风险等级进行识别。该模型具有较高的准确性，但对于已标注的数据量有较高的要求，且模型的普遍适用性和可解释性有待进一步提升。赵长江等[5]基于某市欠税公告数据进行多维关联规则挖掘以发现偷逃税纳税人的特征，为后续税收风险防范提供了有效数据支撑，但该模型也要求有足够的已标注数据才能进行挖掘。胡国庆[6]基于实务工作进行总结，认为当前税收风险识别模型存在指标精准度不高、行业针对性不强、特定复杂事项税收风险识别度低、各税种税收风险识别有效性不一等问题。

综上，当前基于大数据对税收风险的研究大都停留在理论或构想层面，在实际业务中的应用相对薄弱。而聚类作为一种重要的无监督式数据挖掘方法，能够在无先验知识的前提下，结合税收风险管控业务，选择合适的税收风险指标，自主发现税收风险

【基金项目】重庆市教育委员会科学技术项目“大数据背景下考虑行为‘画像’的纳税信用等级动态评估模型研究”(项目编号:KJQN201801103)；重庆市社会科学规划项目“高质量发展下基于大数据的税收政策实施智能化支持机制研究”(项目编号:2018BS68)

疑点。在税收风险疑点发现过程中，聚类不仅可以实现对海量数据的整体分析，而且可以辅助税务人员精确定位税收风险，增加税收风控经验。鉴于此，本文拟提出一种改进的K-means聚类算法并将其用于税收风险疑点识别。基于某地区房地产类企业的股权转让业务验证发现，该方法可以在无先验知识的前提下，更有效地发现异常的企业实例。该模型准确度高，可扩展性强，更具有实用性。

二、改进K-means聚类算法

聚类算法作为无监督学习方法的一种，能够在未标注的实例集中发现实例之间的相似性，并将其分为若干个类。同一类中的实例尽可能相似，不同类中的实例尽可能相异。由此，包含实例较少的小类由于其特征与其他多数实例存在较大的差异，通常被视为可疑实例。聚类的这种特征构成了税收风险疑点发现的理论基础。K-means聚类算法因其典型的基于划分的思想，具有简单易懂、收敛速度快、扩展性强等优势，被广泛应用于各类领域。该算法虽然可以将实例分配到不同的类，但在初始化时不能决定究竟要分几个类以及每个类的中心。因此，使用K-means算法时最好能了解数据的分布，以便确认初始的类别数和质心。然而在税收风险疑点的发现过程中，面对海量高维的企业数据，很难具象化地获取数据的分布情况。这直接影响了聚类的结果和运行时间。

鉴于此，本文针对初始化问题提出一种改进的K-means聚类算法，该方法基于局部的密度信息和全局的相异性信息来确定初始的中心和聚类数目，可以有效提高聚类性能。首先基于实例的最近邻计算各个实例的局部密度，其中密度高的实例被认为更可能成为聚类的中心；然后基于全局的相异性，筛选出彼此相似性最低的实例并将其作为初始聚类的质心；最后基于K-means算法分配实例至各个簇，直至簇中心不再变化为止。具体流程如下：

输入：所有实例，最近邻距离阈值为λ_1，异常阈值为λ_2。输出：各实例所属的类号、类中心以及异常类号。第一步，计算各个实例的局部密度：①计算实例x_i与其他实例之间的距离$d_{ij}(j\neq i)$；②统计d_{ij}中大于等于给定最近邻距离阈值λ_1的数目e_i，将其作为实例x_i的局部密度$\rho_i(i=1,2,\cdots,n)$。第二步，基于全局相异性筛选初始聚类中心：①将局部密度按从大到小的顺序排列，得到序列$sort_\rho_j$，以及相应的实例序列$sort_x_j(j=1,2,\cdots,n)$；②选取局部密度最大的实例作为初始聚类中心之一，即$sort_x_1\in cen$；③$j=2,\cdots,n$，遍历已排序的实例$sort_x_j$，若实例$sort_x_k$既不存在于已选择的类中心的最近邻中，也不与已选择的聚类中心相似，则$sort_x_k\in cen(k\in[2,n])$。第三步，基于选定的初始类中心cen，采用K-means算法进行聚类。第四步，将实例数占总实例数比例小于异常阈值λ_2的类视为异常类。

三、基于改进K-means聚类算法的税收风险识别案例

本文以股权转让中的税收风险识别为例，采用改进的K-means聚类算法对税收风险进行识别。

1. *问题定位、指标选取和数据准备*。股权转让可分为个人股权转让和企业股权转让，其中转让方为个人时，涉及税种为印花税、个人所得税，当转让方为企业时，涉及税种为印花税、企业所得税、契税等。本文就某地区房地产类企业的个人股权转让情况进行分析。根据房地产类企业业务和涉税的特点，拟构建包括财务分析类、税种分析类等27种指标，详见表1。

从工商部门获得某地区2015年427家（其中房地产类企业为23家）企业股权转让的数据，数据包含的主要字段为：统一社会信用代码、注册号、注册资本、生产经营所在区、公司名称、企业类型、股东名称、认缴出资额、认缴出资日期、认缴出资比例、认缴出资方式、住所、主体身份证号码和变更序号等。比对认缴出资金额发现，98%以上的股权变更为平价或低价转让，因此，需要税务部门对变更企业进行税收风险评估，以检测其是否存在不合法的避税行为。

为了保证评估结果的准确性，特从金税三期系统中采集房地产类企业的财务数据和纳税数据作为研究样本。为了保证评估过程的合理性，特提取该区63家房地产类企业2015年1月1日～2015年12月31日的财务报表和纳税数据进行聚类分析。通常企业要按月、季和年填写财务报表，并进行纳税申报。然而，在数据采集时发现部分企业的财务报表项目存在空缺（可能是企业零申报的原因），因此需要根据已有的数据对其进行填充，若缺失的信息太多则只能剔除。最终得到的有效实例数为51。

2. *税收风险疑点分析*。基于财务报表数据和纳税数据计算51家企业的27项税收风险指标，采用改进K-means聚类算法对企业进行分析，聚类结果见表2。

由表2可知，51家房地产类企业共形成了14个

表 1　　税收风险指标

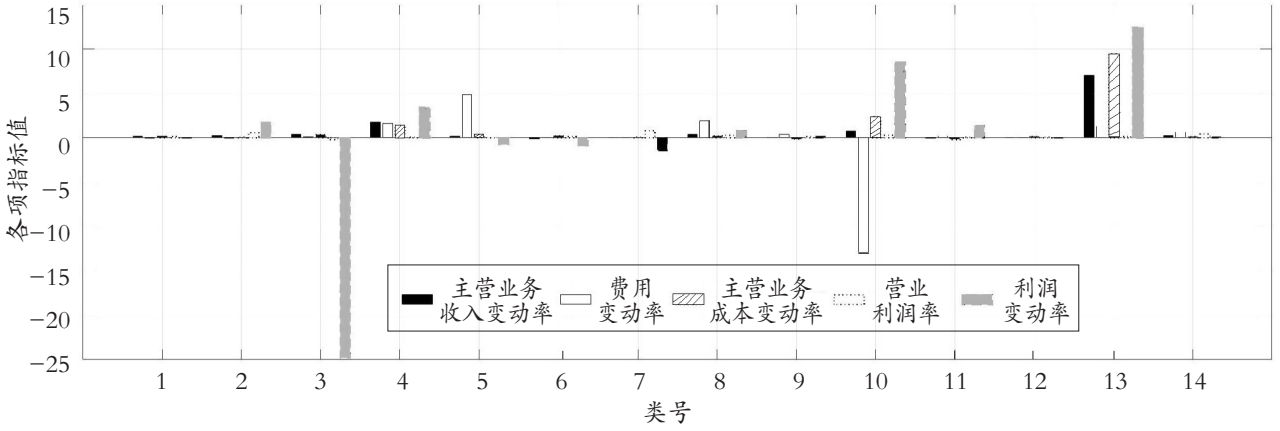
一级指标	二级指标	三级指标	指标描述
财务分析类	收入成本类指标	主营业务收入变动率	(本期主营业务收入-基期主营业务收入)/基期主营业务收入
		主营业务成本变动率	(本期主营业务成本-基期主营业务成本)/基期主营业务成本
	开发费用类指标	营业费用率	营业费用/主营业务收入
		营业费用变动率	(本期营业费用-基期营业费用)/基期营业费用
		管理费用率	管理费用/主营业务收入
		管理费用变动率	(本期管理费用-基期管理费用)/基期管理费用
		财务费用率	财务费用/主营业务收入
		财务费用变动率	(本期财务费用-基期财务费用)/基期财务费用
	利润类指标	营业利润率	利润总额/营业收入
		主营业务利润变动率	(本期主营业务利润-基期主营业务利润)/基期主营业务利润
		应收账款变动率	(期末应收账款-期初应收账款)/期初应收账款
	资产负债类指标	应付账款变动率	(期末应付账款-期初应付账款)/期初应付账款
税种分析类	土地增值税类指标	土地增值税税负变动率	(本期土地增值税负担率-基期土地增值税负担率)/基期土地增值税负担率
		销(预)售收入土地增值税税负率	土地增值税已纳税额/(销售收入+预售收入)
	企业所得税类指标	企业所得税税负变动率	(本期所得税税负率-基期所得税税负率)/基期所得税税负率
		企业所得税税负率	应纳税额/(销售收入+预售收入)
	个人所得税类指标	个人所得税税金工资比	个人所得税/工资支出
		个人所得税权重	个人所得税税额/全部缴纳税
	房产税类指标	从价计征房产税税负率	房产税/房产余值
		从租计征房产税税负率	房产税/租金收入
		房产税税负环比变动系数	本期税负率/上期税负率
	城镇土地使用税类指标	城镇土地使用税税负率	城镇土地使用税已纳税额/计税依据
		城镇土地使用税应纳税额变动率	(本期应纳城镇土地使用税-基期应纳城镇土地使用税)/基期应纳城镇土地使用税
	契税类指标	契税税负率	契税已纳税额/计税依据
		土地交易契税申报计税金额差异率	申报缴纳契税的金额/取得土地成交价格应缴纳契税的金额
	印花税类指标	印花税负担率	印花税已纳税额/计税依据
		印花税应纳税额变动率	(本期应纳印花税-基期应纳印花税)/基期应纳印花税

表 2　　聚类结果

类号	1	2	3	4	5	6	7	8	9	10	11	12	13	14
大小	37	2	1	1	1	1	1	1	1	1	1	1	1	1

类，其中有12个类只包含一个实例，这意味着这12个类的部分指标偏差较大，相应的实例可能存在异常。进一步分析各类的收入变动率、费用变动率、成本变动率、利润率和利润变动率五大特征。其中，费用率是财务费用率、管理费用率和营业费用率的均值，费用变动率是财务费用变动率、管理费用变动率和营业费用变动率的均值。聚类情况如图所示。

由图可知，类1作为大类，其特征表现为五大指标分布比较均衡，这表明在当前环境下，大多数企业的收入、费用、成本和利润变化情况是相对稳定的，可认为该类企业的税收风险较低。类2、6、8、9、10和11这六类的收入、成本、费用和利润变化虽然各不均衡，但是基本匹配，也可断定这六类中企业的税收风险较低。类12和14的收入、费用、成本和利润四项变化幅度都不大，未被归为类1的原因是应付账款或预收账款出现大幅上涨(类12为18.54倍，类14为1.42倍)，可能存在虚构专票、未及时确认收入等涉税问题。此外，类4的收入、成本和费用都出现了一定幅度的增加，利润也相应地上涨。与之相类似，类13的收入、成本和费用都出现了大幅的增加，利润也相应地大幅上涨。不同的是，类4的应付账款上涨



聚类情况概览图

了36.9倍，可能存在虚构专票等涉税问题；类13的应收账款短期内上涨了14.88倍，可能存在对外虚开票据、对外融资等涉税问题。

值得一提的是，类3的利润出现大幅下滑，而收入和成本、费用相对变化不大，与利润变化不相匹配，税收疑点很明显。类5的费用大幅上涨，利润下降，可能存在多计费用、少计收入的税收风险。类7的利润出现一定幅度的下滑，而收入和成本、费用几乎没有变化，与利润变化不相匹配，税收疑点明显。

3. 税收风险疑点验证。由上述分析可知，类3、5和7中的企业（对应第9、14和21号企业）存在明显的税收风险，其中第9类中的企业在2015年发生了股权转让，并且其转让形式为平价转让，转让情况见表3。

表3　　变更前后股权占比情况

股东	变更前股权份额	变更后股权份额	变更时间
自然人股东A	100%	0	2015
自然人股东B	0%	100%	2015

该企业创立于2007年，注册资金为16亿元。2015年自然人股东A将其全部股份转让给自然人股东B。税务人员通过爬取并分析企业官网的相关信息，基于聚类结果并结合初步的取证分析，发现该企业在股权转让第一环节净资产评估中存在明显税收风险；基于网上的公司介绍，粗略估计其实际总资产在2014年就已上涨了10.19倍，所有者权益达近9亿元。因此，2015年股权平价转让形式不合理。为此，税收工作人员多次约谈企业负责人和相关财务人员，并进一步调查和精确评估了其股权交易时的企业净资产，测算其应补缴个人所得税近5千万元。

值得注意的是，基于聚类的方式挖掘出的小类并不一定都存在问题，需要税务人员对可疑企业进行进一步分析排查。聚类结果作为一种导向，可帮助税务人员快速定位可疑企业，缩小排查范围。

四、结语

本文以房地产类企业的财税数据为实验样本，结合网络爬取数据，验证了改进K-means聚类方法在税收疑点发现上的有效性。基于改进K-means聚类方法的税收风险识别兼顾了对大数据的总体分析，可发现与总体差异较大的异常实例，有效地提高了税务风险监控效率。虽然该方法下税务人员不需要先验知识就可进行风险识别，但在判定企业是否存在高风险时仍需要与其经验判断相结合。

主要参考文献：

[1] 朱丹.“金税三期”背后的税收风险管理探讨[J].现代商贸工业，2018(20)：109～110.

[2] 徐壁.基于大数据的税收风险管理研究与应用[J].信息与电脑(理论版)，2018(23)：102～103.

[3] 刘小瑜，温有栋，江炳官.“互联网+”背景下高新技术企业的税收风险预警——基于智能优化算法的研究[J].税务研究，2018(6)：82～88.

[4] 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(9)：19～26.

[5] 赵长江，吴乐云.多维关联规则挖掘在欠税管理中的应用[J].科技广场，2015(12)：29～33.

[6] 胡国庆.税收风险识别模型建设存在的问题及对策[J].现代经济信息，2016(23)：173～174.

作者单位：1.重庆理工大学会计学院，重庆 400054；2.重庆市渝北区税务局，重庆 401120