供应链金融大数据分布特征的分析与洞见

刘 颖
(吉林财经大学管理科学与信息工程学院 长春 130117)
(吉林省物流产业经济与智能物流重点实验室 长春 130117)
(吉林财经大学互联网金融重点实验室 长春 130117)

摘 要 半结构、非结构化、海量的供应链金融数据使得大数据环境下金融数据分析的模式和方法相对复杂。面向大数据样本研究,如何将大样本相比于小样本的独有特征体现在分类模型中值得深入探索。文中从供应链金融数据分布特征入手,分析影响信用风险分类模型的主要因素;对多年来的相关研究成果进行归类分析,概括信用数据分布特征,包括信用数据非均衡与不对称性、信用数据噪声和离群点的存在以及信用数据的非线性多维特征,并探讨了进一步的解决策略。供应链金融大数据分布特征的分析旨在助力挖掘隐含在海量金融数据背后的知识信息,为信用风险模型的构建奠定了坚实的基础。

关键词 供应链金融,信用风险,大数据,分布特征,非均衡数据,离群点,多维
中图法分类号 TP399 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2019.02.001

Big Data Analytics and Insights in Distribution Characteristics of Supply Chain Finance

LIU Ying
(School of Management Science and Information Engineering, Jilin University of Finance and Economics, Changchun 130117, China)
(Jilin Province Key Laboratory of Logistics Industry Economy and Intelligent Logistics, Changchun 130117, China)
(Laboratory of Internet Finance, Jilin University of Finance and Economics, Changchun 130117, China)

Abstract The semi-structured, unstructured and massive supply chain finance data make the analysis method relatively complicated in large data environment. How to use the unique characteristics of large samples to improve classification performance is worth exploring for the research on large data samples. This paper analyzed the main factors, which affect the classification model of credit risk based on the distribution characteristics of financial data in supply chain, proposed distribution characteristics of credit data after researching the relevant achievements over the years, including imbalance data, noise and outliers, nonlinear multidimensional and so on, and then discussed further solutions to mine the knowledge of the massive financial data, which provides an effective theoretical basis for the construction of credit risk model.

Keywords Supply chain finance, Credit risk, Big data, Distribution characteristics, Imbalance data, Outliers, Multi-dimension

1 引言

大数据(Big Data,BD)是时代话题,也是时代机遇。在当今数字时代,源于多重媒介的海量数据将继续大幅增长,国际数据公司 IDC(International Data Corporation)报告指出,全球创造和复制的数据总量为 1.8 ZB(1021 B),它在五年内增加了近 9 倍[1-2]。大量增加的数据来源于工业、企业供应链(SC)、网络外围使用的各种设备,包括嵌入式传感器、智能手机、计算机系统等[3]。大数据时代的到来使得原本杂乱无章的数据成为探究用户行为规律、挖掘商业价值和社会效益的媒介;同时,它也为获取更多的知识信息创造了新的机会。但"数据丰富,信息贫乏(Data Rich & Information Poor)"现象不可避免。久而久之,未被深度挖掘的数据可能成为"食之无肉,弃之可惜"的鸡肋。Gantz 等[4]描述了大数据 5V 理论,即体积(Volume)、变化(Variety)、速度(Velocity)、准确性(Veracity)和价值(Value)。在 5V 理论中,大数据的准确性和价值尤其重要,没有数据分析,其他 BD 处理方面(如选择、存储和管理)将不会产生更多价值[5-8]。归根结底,真正的大数据应用体现在数据挖掘的深度,其价值实现还是依赖于数据挖掘技术[9]。

供应链金融是供应链管理发展到一定阶段的必然产物,亦是供应链管理学和金融管理学两个学科的共同产物。近年

到稿日期:2018-08-30 返修日期:2018-11-01 本文受吉林省科技厅自然基金(20180101337JC),国家自然科学基金(61402193,61806082),长春市地院(校、所)合作项目(17DY009),物流产业经济与智能物流吉林省高校重点实验室开放基金(201702)资助。
刘 颖(1979—),女,博士,副教授,CCF 会员,主要研究方向为数据挖掘、金融工程,E-mail:lyaihua1995@163.com(通信作者)。

来,信贷机构在构建信用风险评估模型时能够利用的样本数量与信息规模都急剧增加,信用风险管理也逐渐进入“大样本”时代。面向大数据样本研究,学者们多数只是直接将现有基于小样本设计的模型应用在大样本上,大样本相比于小样本的独有特征并未在模型的设计中得到体现。大样本数据集是否存在某种分布特征?能否根据该特征设计出性能更为优越的信用评估模型?这些都是值得研究的重点。本文从供应链金融大数据分布特征入手,以金融数据分析为切入点,分析影响信用风险分类模型的主要因素;对多年来的相关研究成果进行归类分析,概括信用数据分布的非均衡与不对称性、信用数据的噪声和离群点的存在以及非线性多维等特征,并探讨了进一步的解决策略,旨在获取隐含在海量金融数据背后的知识信息。

2 供应链金融大数据分析的现状

经济全球化发展、金融衍生品的加速膨胀、银行同业的竞争白热化及其他一系列复杂的相关原因,导致全球金融市场波动加剧,客户违约行为不断出现。在此背景下,金融业对风险控制提出了更高的要求,由此也引发了信用风险的计量和管理方法的革命性变化。供应链是以核心企业为中心,通过管理物资流、信息流和现金流,将核心企业上游各级供应商、下游各级分销商以及终端客户连接成一个整体的网络结构。供应链管理负责供应链上成员之间的管理和控制,增强供应链整体的竞争实力,提高链上成员的工作效率和收益,使得所有链上的成员成为一个协调发展、不可分割的整体[10]。但实际上,业界及学界对于供应链管理关注的重点一直偏向于物资流和信息流管理的整合,忽视了对资金流的综合控制。供应链金融是指银行等金融或物流机构在供应链运作的全过程中,以核心企业为出发点,向中小企业客户提供结算和融资的服务。供应链金融作为新兴信贷服务模式,成为商业银行业务新的重要增长点,其实质是为了解决相对弱势的中小企业的融资问题。2014 年,金融咨询及供应链融资电子平台提供商 Demica 公司发布的一份报告显示:2011－2013 年期间,国际银行的供应链金融业务增长率为 30%～40%;预计到 2020 年,供应链金融业务的年增长速率至少达到 10%[11]。供应链金融在缓解中小企业资金缺乏、松绑核心企业的资金约束及促进供应链稳定发展方面可望起到积极的作用。

与传统融资模式相比,供应链融资包含的资金关系错综复杂,评估指标动态多样,使得供应链融资的风险及风险评价有其自身的特征及难度。世界银行对全球银行业危机的研究表明,导致银行破产的主要原因就是信用风险[12]。除去企业本身的短期债务比率、流动资金、财务杠杆比率等因素,客户质量和违约状况仍是授信银行的主要风险来源[13]。通过有效的风险评价方法来深度挖掘融资企业的数据,准确、客观、公正地评价企业的资信状况,可为大量风险企业提供及时的资金融通,最终实现融资企业与银行间的信息共享和合作共赢。

大数据分析(Big Data Analysis,BDA)是一个利用正确的、可解释的形式从数据中提取知识的技术驱动生态系统。BDA 借助 NoSQL,BigQuery,Map Reduce,Hadoop,Flume,Mahout,Spark,WibiData 和 Skytree 之类的高级工具,为商业智能分析[14]、医疗保健分析[15]、社会媒体分析[16]、智能城市[17]、智能运输管理[18]、金融与会计[19]、金融风险管理[20]等各个领域提供智能决策。为了较好地洞察目前供应链金融信用风险评价领域的相关研究现状,本文以 web of science 数据库为数字平台,索引 2013－2018 年间的期刊、评论、会议论文、会议摘要等多种成果形式,搜索关键词包括“供应链金融”“风险”“信用风险”,并以主题、标题、主题与标题兼而有之的形式进行检索。具体查询抽取过程如表 1 所列;查询结果如表 2 所列,其中包含供应链金融研究热点的前 10 名期刊的名称及研究数量。

表 1 查询抽取过程的实例

Table 1 Examples of query extracting process

问题	形式	关键词	缩写
Q1	主题	供应链金融	To-Q1
Q2	主题	供应链金融 且 风险	To-Q2
Q3	主题	供应链金融 且 信用风险	To-Q3
Q1	标题	供应链金融	Ti-Q1
Q2	标题	供应链金融 且 风险	Ti-Q2
Q3	标题	供应链金融 且 信用风险	Ti-Q3
Q1	主题＋标题	供应链金融(主题)且 风险(标题)	ToTi-Q1
Q2	标题＋主题	供应链金融(标题)且 风险(主题)	ToTi-Q2

表 2 前 10 个供应链金融风险研究领域期刊

Table 2 Top 10 research area journals of SCF contributions

期刊名	主题			标题			主题和标题	
	To-Q1	To-Q2	To-Q3	Ti-Q1	Ti-Q2	Ti-Q3	ToTi-Q1	ToTi-Q2
OPERATIONS RESEARCH MANAGEMENT SCIENCE	100	49	32	15	3	1	17	10
MANAGEMENT	79	33	17	12	1	0	12	6
ENGINEERING INDUSTRIAL	48	21	15	8	0	0	4	5
ENGINEERING MANUFACTURING	42	19	10	4	0	0	5	1
ENVIRONMENTAL SCIENCES	30	2	1	4	1	1	1	2
ECONOMICS	29	9	3	3	1	0	3	3
ENVIRONMENTAL STUDIES	21	5	2	4	1	1	2	2
MATHEMATICS INTERDISCIPLINARY APPLICATIONS	20	7	5	11	3	1	4	5
GREEN SUSTAINABLE SCIENCE TECHNOLOGY	18	2	1	4	1	1	1	2
BUSINESS	15	7	3	2	1	0	3	2

图 1 显示出该领域近 6 年的研究趋势,从研究总体数量上看,论文数量稳中有升,2018 年论文的总量约为 2013 年论文总量的 2 倍。

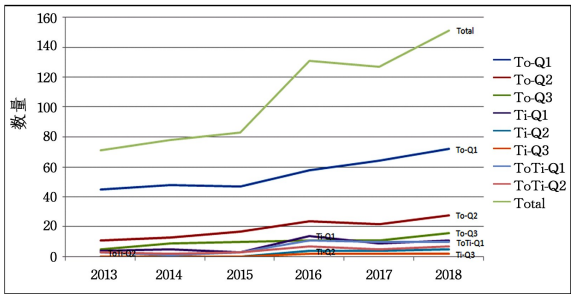

图 1　近 6 年 WOS 文章数量的分布

Fig. 1　Distribution of articles in past 6 years in WOS

3　供应链金融信用数据的分布特征

半结构、非结构化、海量的供应链金融数据使得大数据环境下数据分析的模式和方法相对复杂。通过大数据应用分析有效地增加价值(增值),能够使运营或供应链管理发挥优势。本文对相关研究成果进行归类分析,并概括影响信用风险评价的主要数据分布特征,包括信用数据的非均衡性与不对称性、信用数据的噪声和离群点的存在以及信用数据的非线性多维特征。

3.1　信用数据的非均衡性与不对称性

信用评估的实质是一个二元客户分类问题。信用样本的获取具有涌现性,即大量样本中的有用样本点往往很少。这种某类样本明显少于其他类别样本的样本集合被称为不均衡样本。概括地说,信用评估问题具有薄靶、类别分布不平衡与不对称等特性[12]。数据非均衡与不对称对供应链金融风险评价形成了较大制约,将一个信用差的客户错误分类给企业造成的损失往往比误分一个信用好的客户带来的损失大得多。

3.2　信用数据噪声和离群点的存在

在信用风险评价领域,噪声离群点包括人为财务造假所带来的信息失真、周期性统计误差和报告偏倚所产生的错误。噪音离群点也叫孤立点、异常点等,离群点的存在对分类精度的影响较为明显。通常,引起离群的原因包括数据来源于异类、数据变量的固有变化,或者在数据测量和收集阶段产生的误差。离群点的消除具有重要意义,如果进行简单的剔除操作,可能会丢失有意义的信息。处理好高维空间中数据的稀疏问题,合理解释离群点形成的原因,选择合适的度量方法,具有重要意义。

3.3　信用数据的非线性多维特征

信用风险评估特征具有高维、非线性、动态等特点。现有的分类方法大多是基于数据之间的相似度来实现的,而在高维情况下,数据十分稀疏,数据点之间的距离及区域密度不再具有直观的意义。此外,高维度的数据对数据样本数量的要求更高,当数据维度远大于样本数目时,极易造成过拟合问题,因此寻找高维数据的本质低维结构,即解决数据降维问题十分必要。

4　数据分布特征的相关解决策略

4.1　非均衡样本的解决策略

不平衡分类问题是指在一个分类问题中某些类的样本数量远多于其他类别的样本数量[21]。在大多数决策过程中,每种可供选择的方案的适用范围都存在不确定性。在处理不均衡数据集时,分类器将倾向于多数样本类,从而导致错分类。针对以上问题,解决非均衡问题主要集中在以下几个方面。

(1)基于数据分布角度

数据分布的调整主要从数据准备阶段入手,通过数据重采样或数据分组等手段使得类别在一定程度上达到平衡,从而消除类别不平衡问题[21]。学者们多采用重抽样方法(例如随机向上抽样方法和随机向下抽样方法)来平衡类别分布,然后训练分类模型。通常,利用重抽样算法扩充少数样本数量会由于样本太少而造成过拟合问题,难以取得满意的效果。Debashree 等[22]提出了一种基于密集最近邻和 Tomeklink 欠采样技术的改进算法,该算法通过检测离群点、删除多数类中的噪声及冗余样本来平衡数据集。Yang 等[23]基于数据结构挖掘不平衡数据之间的内在关系,从多类样本的隐藏数据结构角度出发,提出利用多类样本子集群的结构找出两类样本的内在联系,指出样本不平衡包括类内不平衡和类间不平衡,并通过发现不平衡样本分类中多数类的隐藏结构来改善分类器性能。模型在第一阶段利用聚类算法学习多数类样本的隐藏结构;在第二阶段调整少数类样本的分类边界。通过实验可以看出,所提模型能较好地捕获隐藏的数据结构,改变非均衡样本的分类性能。衣柏衡等[24]针对传统 Smote 算法在处理非均衡数据时产生过拟合的问题,提出仅对错分样本进行人工合成的改进思想,即仅选择上一次迭代中被错分的样本作为下一次迭代的起始样本,直到少数类样本和多数类样本的数量均衡或不再有少数类样本被错分。Pierri 等[25]同时考虑 3 种方法来解决非均衡信用数据的分级问题,分别是案例控制匹配的 Logistic 回归、平衡样本的 Logistic 回归和 ROSE(Random Over Sampling)平衡样本回归,并将其用于解决中小企业的信用评级问题。

(2)基于监督学习策略

随着机器学习研究的深入,神经网络算法、支持向量机(Support Vector Machines,SVM)等监督学习策略为解决非均衡样本提供了新的思路。Li 等[26]提出了一种新的基于子空间与贝叶斯神经网络集成的深度方差网络算法,相比于传统的神经网络算法,该算法对非均衡样本分类能产生更好的识别效果。所提算法主要考虑样本数据的类内和类间的异构性,将贝叶斯模型引入到传统神经网络学习框架中,将在每个训练周期中得到的特征聚类到与判别相关的子空间中,以指导在非平衡训练数据集上的同质性和异质性的学习及自适应调整。熊冰妍等[27]提出了一种基于样本权重的欠采样方法

KacBag(K-means AdaCost Bagging),该算法以 Bagging 算法为框架,使用 K-means 算法对数据集进行多次聚类,并根据聚类结果更新样本权重,找出位于多数类中心区域的样本,然后通过样本权重对多数类进行欠采样,并与少数类样本组成多个平衡数据集,在多个平衡数据集上应用决策树算法得到若干个弱分类器,最后通过弱分类器加权投票来生成最终的分类器。Chiclana 等[28]基于语义角度提出 Type-1 型有序加权平均算子(T1OWA)的模糊方法来解决不平衡数据问题,并以 T1OWA 算子法为基础,采用不平衡模糊语言标签对欧洲债券的信用价值进行评估。神经网络方法虽然克服了假设限制,但也存在陷入局部极值、出现"过拟合"等缺点。支持向量机由 Vpanik[29]于 1995 年提出后便受到广泛关注。基于结构风险最小化的 SVM 弥补了传统分类器的"过拟合"及小样本等不足,但 SVM 对非均衡样本分类时分类超平面会向少数类偏移,从而导致将更多的少数类样本错分为多数类。Shao 等[30]提出了一种基于不同训练点的加权拉格朗日双支持向量机(WLTSVM)算法,该算法构造两个近似超平面来解决非平衡数据分类问题,分类性能大幅提高。程砚秋[31]指出现有信用风险评价方法较少考虑违约样本和非违约样本不均衡对信用风险评价模型产生的影响,并根据样本差距与权重关系成正比的思路提出了不均衡 SVM 分类模型。该模型分析了有无特定评价指标、特定评价指标的数值变化对贷款客户违约状况的影响,确定了信用风险评价指标的组合权重,进而构建了具有显著区分能力的评价方程,以违约样本正确识别率、违约样本的准确率与查全率等因素作为贷款客户违约状况识别率的度量标准,改变了样本数据不均衡所导致的样本总体精度很高、违约样本正确率反而不高的现象。

(3)基于半监督学习策略

不难看出,上述方法都属于监督式分类建模的研究范式,即事先假设已经存在相当数量的标签样本用于建立风险评级模型。然而,供应链融资的复杂性使得大量准确的标签样本难以获得,此时,仅依靠少量的标签数据无法准确度量实际的数据分布规律。半监督学习(Semi-supervised Learning)因近年来随着机器学习技术的不断发展,利用无标签样本这一需求越来越强烈而受到广泛关注。半监督学习所需样本既包括已知类别样本,也包括未知类别样本,通过挖掘未知类别样本中所蕴含的固有结构信息,对拟合分类器进行校正[32]。夏战国等[33]提出类不均衡的半监督高斯过程分类算法,首先依据类不均衡数据特性对数据进行预处理;然后用少量的标记数据进行高斯过程分类训练,选取预测概率置信度最高的未标记数据加入到原有的训练集中,用扩充后的训练集再次进行高斯过程分类;最后通过自训练迭代执行构造出最优的高斯过程分类器,用来对测试数据集进行分类。肖进等[12]提出类别不平衡环境下基于随机子空间(RSS)的半监督模型,该模型首先利用 RSS 方法在模型训练集上得到若干基本分类器,从大量无类别标签数据集中选择性地标记一部分最合适的样本加入到原始训练集中;在最终的训练集上训练分类模型,并对测试集样本进行分类。在 3 个客户信用评估数据集上进行的实证分析表明,RSSCI 模型的信用评估性能不仅优于常用的监督式集成信用评估模型,还优于已有的一些半监督协同训练信用评估模型。

(4)基于集成学习策略

单一分类模型很难实现在整个样本空间上的准确分类,而如果将多个分类器的分类结果进行集成,让每一个分类模型都在其优势空间区域内发挥作用,将有望提高客户信用评估模型的准确性[12]。集成学习策略应运而生,其主要思想是通过多个弱分类器的组合生成一个强分类器。集成学习中的 Boosting 算法简单有效,不少学者用它来处理不平衡数据集的分类问题。李雄飞等[34]将过采样技术与 Boosting 相结合,在每一次迭代中增加合成的少数类样本,并及时删除被误分的合成样本,以防止产生错分样本而影响算法性能,最后利用决策树算法训练多个弱分类器,并将多个弱分类器集成为最终的分类器。李克文等[35]提出了一种将采样技术与 Boosting 算法相结合的分类算法 PSBoost,该算法首先应用 Smote 算法对少数类实现过采样,然后在不改变数据分布的情况下对所有数据随机欠采样,最后再与 AdaBooting 算法相结合来完成数据分类。朱兵等[36]以基于集成技术的迁移装袋模型(TrBagg)为基础,使用两阶段抽样获取用于学习基模型的训练集以更好地对少数类样本进行分类,同时使用数据分组处理技术(GMDH)作为其基模型的集成策略。Chang 等[37]从时间角度考虑风险评价模型的构建,提出一种基于决策树的短期信用风险评估模型。其目标是使用决策树来过滤短期违约,以产生可以区分违约贷款的高精度模型。该模型将通过融合(Bagging)集成算法和对少数类样本采用过采样技术(Smote),来提高决策树的稳定性和划分非平衡数据的性能。Sun 等[38]指出,在信用风险评价过程中,特征选择和非均衡数据的处理至关重要,同时提出以 T-test 和分支定界(B&B)为基础的动态特征选择模型,并以 SVM 和多重判别分析为基分类器进行集成以处理非均衡样本模型(IOMCE)。实验分别将单分类器+非特征选择模型、单分类器+5 个不同特征选择模型与 IOMCE 进行比较,结果表明 IOMCE 模型能有效处理非均衡信用样本的分类问题。同时,利用特征选择模型降低分类数据的维度对提高非均衡信用样本的评价精度也至关重要。

4.2 噪声离群点的解决策略

目前,基于噪声离群点剔除方法包括基于统计学的探测方法、基于聚类的探测方法、基于距离的探测方法,具体研究情况如下。

(1)基于统计学的探测方法

假设检验是最早用来发现异常样本点的统计学方法。通常,基于统计学的离群点检测方法必须假设数据集符合特定的分布模型,如泊松分布或正态分布等。基于正态分布的一个具有代表性的离群点被定义为偏离平均值 μ 超过 3σ 的数据点,但"3σ 准则"的均值和标准差对离群点十分敏感,对此

学者们陆续提出基于模糊自回归隐马尔可夫模型的控制过程数据离群点检测方法[39]。Garces 等[40]则通过建立输入与输出数据的非线性回归模型,并分析测量值与预测值之间的偏差,实现对过程数据离群点的检测。上述方法适用于连续过程数据,而间歇过程变量的轨迹通常随着时间呈现出更为复杂的非线性变化趋势。为此,贾润达等[41]提出了一种基于鲁棒 M 估计的间歇过程离群点检测方法。通过积分方程离散化将参数估计问题转化为最小二乘优化问题,并利用 Tikhonov 正则化方法及鲁棒 M 估计消除噪声及离群点对参数估计的影响,同时通过分析各个样本点的权值实现对离群点的检测。姜震等[42]利用高斯模型表示样本分布,提出集成自学习算法。该算法利用迭代训练中生成的分类器构成的集合预测样本标签,将自训练方法和集成学习相结合来提高无标签样本标注的准确率,根据所提噪声量化方法评估当前噪声下分类器的性能,并据此及时移除有可能造成分类性能下降的伪标签样本,从而实现迭代训练的可回溯机制。吴建华等[43]利用信息干扰噪声的偏倚性来捕获财务报告信息中资产价值的潜在偏倚,推导了信息偏误下资产价值的条件分布、违约概率和信用价差解析表达式。研究发现,在不确定性的市场环境中,创造性的财务和自由裁量权所导致的向上偏倚会导致噪声与违约概率之间形成一种非线性的同向关系,从而进一步造成违约强度的上升。反之,财务保守造成的向下偏倚则会导致噪声与违约概率之间形成一种反向关系。该模型为理解债券违约风险和财务报告信息扭曲之间的关系提供了新的启发,可以帮助债券投资者在信息披露问题比较严重的市场环境中评估违约风险。

概率统计模型探查离群点只需存储描述模型的最少量的信息。该方法的使用前提是假设数据符合某种分布规律且基于单个属性,不适合用于解决高维数据离群点探测问题。

(2)基于聚类的探测方法

基于聚类的方法是通过考察对象与簇之间的关系,首先将数据集分成若干类簇,如果一个对象不属于任何类簇,则该对象被认为是基于聚类的离群点。Jiang 等[44]提出的两阶段聚类离群检测算法是这类算法中的典型代表。该算法的第一阶段利用改进的 K-means 算法将整个数据集划分成若干聚类;第二阶段将形成的各个簇类用其质心代替,形成新的数据集,并以质心之间的距离作为权值生成一棵最小生成树,然后删除树中的长边,从而形成多棵子树。那些具有较少结点的树所对应的小簇类被认为是离群点或离群簇。Zhuang 等[45]提出了在异质网中根据离群性排序子网中节点的离群点检测方法 BMSim,首先在各个子网中排序网络中的节点以查找出离群点,进而分析整个网络中的离群点。朱利等[46]主要针对复杂分布的数据集中具有多种类型的离群点的问题,融合基于密度和基于聚类方法的优势,提出了一种新的数据结构来关联数据点之间的信息,并在此基础上提高算法中数据点近邻的查询效率,设计了新的离群因子公式,自适应地判断簇类数目,从而有效检测离群点。彭涛等[47]通过研究一种双类型离群点检测方法,提出排序和聚类相结合的方法来提高离群点检测的效率;将属性对象作为目标对象的特征表示,对目标对象进行聚类,并通过分析属性对象的数据分布来检测离群点。刘颖等[48]提出一种基于离群点剔除的支持向量机信用风险评价模型,该模型利用模糊聚类算法删除距离聚类中心较远的样本点,避免因奇异点的存在而导致的分类精度较低的问题;他们同时还比较了 SVM 核函数参数和关键特征对分类模型的影响程度。

基于聚类的探测方法属于无监督方式,对许多数据类型均有效。但是,聚类算法的主要目的是类簇的探测,而非发现离群点,因此离群点检测的效率较低。同时,不同聚类算法适合特定的数据类型,算法的针对性很强。而对于海量数据集,聚类方法的开销通常很大,这可能是一个瓶颈。

(3)基于距离的探测方法

为了弥补以上算法的不足,学者们提出了基于距离的离群点剔除算法。Knorr 等[49]给出了基于距离的离群点的定义:给定两个参数 k 和 r,对于数据集中任意的数据点 p,若与 p 的距离不大于 r 的点的数目少于 k,那么 p 为离群点。

针对基于距离的离群点,Wang 等[50]和 Pillutla 等[51]分别就不同的离群点定义,提出了基于局部敏感哈希的离群检测算法。王习特等[52]提出一种新型的分布式计算方法,首先利用 BDSP(Balance Driven Spatial Partitioning)空间数据划分方法对数据进行预处理,并基于 BDSP 算法引出一种 BOD(BDSP-based Outlier Detection)离群点检测算法。在每个计算节点本地,该算法使用 R 树索引进行批量过滤,快速计算出本地离群点并得到候选集;进一步地,使用 BDSP 中提供的块编码规则确定需要通信的相邻块,并计算出最终结果。由于离散型属性值之间并没有类似于连续型属性值之间那样固有的距离度量关系,所以将基于距离的方法直接应用到包含离散型属性的数据集上是不合适的。针对这一问题,江峰等[53]引入粗糙集理论,利用粗糙集解决离散型属性的处理问题,提出粗糙集中的重叠度量、值差异度量和加权重叠度量 3 种面向离散型属性的距离度量方法。姚潇等[54]为提高支持向量机对信用风险评估的精度,引入模糊隶属度对近似支持向量机进行改进,以保留近似支持向量机泛化能力强、容易求解的优点,同时消除噪声样本对近似支持向量机模型的影响。刘京礼等[55]针对传统最小二乘支持向量机对噪声点不敏感的特点,提出了一个鲁棒赋权自适应 Lp 最小二乘支持向量机模型,该模型能够适应信用评估样本数据库类别不均衡的特点,可以有效处理信用评估数据中带有噪声点的问题。将该模型应用于仿真数据集和 3 个消费者信用风险评估实例中,实验结果表明,所提模型具有较好的鲁棒性,能够适应信用数据库类别不均衡的特点。也有学者提出分块检测数据集的 iORCA 算法[56],即在数据集中随机选取支撑点,然后计算所有对象与支撑点的距离并按降序建立索引。为避免将离群点选取为支撑点,许红龙等[9]明确了支撑点选取的两大目标,即边缘支撑点和密集支撑点,并提出基于多种支撑点的度量

空间离群检测算法 VPOD。在近似的密集区域选取支撑点，即密集支撑点，对应使用终止规则，用 FFT(Farthest-First Traversal)算法另选取若干支撑点，即边缘支撑点，与数据集计算距离而形成支撑点空间，并利用距离三角不等性，使距离的计算次数显著减少，从而提高检测速度。

基于距离的离群点检测克服了基于统计的离群点检测算法中对数据的分布模型和分布参数的限制，并且适用于任何维度的特征空间。然而，基于距离的离群点检测方法需要大量的距离计算，而且不能检测出局部离群点。

(4)基于密度的探测方法

以上离群点的定义与剔除均是基于全局观察的，但许多实际数据集的结构更复杂，因而还存在另一种相对于全局观察的数据异常，即局部离群点。Breunig[57]提出基于密度的离群点检测算法，其根据数据点的邻域信息考察数据点与其他近邻“密度”的差异性来判断该数据点是否是离群点。这种差异性被称为局部离群因子(Local Outlier Factor，LOF)。Jin 等[58]基于逆近邻(Reverse Nearest Neighborhood，RNN)提出离群影响因子(Influenced Outlierness，INFLO)，使局部离群点的挖掘效果进一步得到提高。现有的局部离群点检测算法一般不分区数据对象，使得计算的复杂度较高。对此，周世波等[59]提出了一种基于偏离的局部离群点检测算法，该算法通过对数据集进行数据块的划分，并计算每个数据块中数据对象的离散程度，来解决局部离群对象的度量标准问题，从而发现可能存在的局部离群点。

4.3 非线性多维特征的解决策略

目前常用的特征降维方法大致分为线性与非线性方法[60]。通过线性降维方法得到的低维数据仍可保持高维数据之间的线性关系。线性方法主要包括主成分分析法(PCA)、线性判别分析法(LDA)、局部保留投影法(LPP)等。当数据位于高维空间中的一个低维线性超平面上时，此类方法能够有效对其实行降维。线性降维方法不能对具有低维非线性分布结构的高维数据进行有效的降维处理，这极大限制了它们的应用范围[61]。近年来出现了一类新型的非线性降维方法，针对高度非线性结构的数据集合，非线性降维方法更能揭示数据的内在特征。此类方法的典型代表包括等距特征映射(Isometric Mapping，Isomap)、局部线性嵌入(Locally Linear Embedding，LLE)与拉普拉斯特征映射(Laplacian Eigenmap，LE)等。非线性降维算法的计算性能对数据的非线性流形结构具有自适应性，只涉及到较少的参数选择问题；另外，基于非常易于理解的模型构造方式，降维后的数据特征具有很好的可解释性。目前，这类方法主要被应用在聚类与数据可视化等领域中，在模式分类、回归分析等需要预测功能的数据挖掘方面会失效。

性信用评估根据待评估样本的多维度属特征，设计合适的模型将所有样本分为若干类[62]。应用于信用风险评价的降维技术主性要包括线性降维方法、特征选择方法、投影变换方法等。

(1)线性降维方法

通过线性降维方法得到的低维数据仍可保持高维数据之间的线性关系。线性方法主要包括主成分分析法(PCA)、线性判别分析法(LDA)等。Chen 等[63]以中国台湾公司为研究样本，采用主成分分析首先对原始数据进行预处理，然后在此基础上构建以神经网络和数据挖掘为基础的财务危机预测模型。潘和平等[64]提出一种将 EMD，PCA，ANN 方法融合的金融时间序列自适应预测模型，该模型基于分解-择优-综合的信息融合思想，利用 PCA 算法进行降维，并将降维后的几个主成分输入神经网络以实现组合预测。West[65]为了研究商业银行信用评价的准确性，利用判别分析法建立了信用评价模型，并分别对德国和澳大利亚两组财务数据进行两类模式的分类，判别准确率分别达到 72.6%和 82.96%。基于线性技术的 PCA 方法对于过程误差、噪声、冗余信息以及变量之间的非线性数据的表现较差。而通过对信用风险实证的研究发现，企业信用风险的高低与企业财务比率之间的关系以及各财务比率之间的相互关系在绝大多数情况下均呈非线性关系。此时，若用线性 PCA 方法进行处理，则会丢失大量的信息，势必导致数据的实际特征提取存在偏差，甚至出现错误[66]。熊志斌[67]借鉴传统方法中的序数主成分概念，提出基于顺序自联想神经网络(SAANN)的非线性主成分分析法(NLPCA)。结合神经网络(NN)和 Logisitic 模型，以上市公司为研究对象，分别构建基于 NLPCA-NN 和 NLPCA-logisitic 的信用评估模型。张洪祥等[68]应用多维时间序列数据对受评样本进行信用评价，由于多维时间序列数据的信息量一般较大，同时维数增多容易引起“维数灾难”，他们按照采样时间点对多维时间序列数据进行分割处理，在每一个独立采样时间点上进行灰色关联分析，并将得到的受评样本在各个时间采样点上的灰色关联度值作为信用评判值，从而达到降维的目的。张娟等[69]指出信用数据中刻画违约概率的预测变量类型丰富，变量维数较高，并且变量之间经常存在共线性等，利用广义半参数可加模型对客户违约概率进行建模，并将 Group Lasso 方法应用于模型进行变量选择和估计。

(2)投影变换方法

Isomap 是由 Tenenbaum 等提出的一种非线性流行学习算法[70]。李菲雅等[71]考虑到财务数据特征的非线性和高维性，采用等距特征映射(Isomap)算法对财务指标进行特征提取，以减少数据的冗余。Lin 等[72]分别利用 Isomap-SVM，PCA-SVM 和 SVM 方法构建风险评价模型，其中 Isomap-SVM 具有较高的分类精度，并且实验结果表明其在处理非线性数据分类问题时能较好地降低特征维度。Ribeiro 等[73]利用 Isomap 和 SVM 构造了半监督的 Isomap 模型，分别利用半监督 Isomap 模型、SVM、RVM 和 KNN 对 1000 多家法国工业企业进行了破产预测，结果表明半监督 Isomap 模型的分类精度较高。Tong 等[74]引入 Isomap 用于属性降维，并将相关向量机(RVM)用于信用分类，构建了一个 Isomap-RVM 模型来对中国上市公司进行财务分析，并将该模型与 Isomap-

SVM,PCA-RVM 进行比较,结果显示所提模型的分类效果较好。投影变换的主要局限在于低维子空间的重叠和数据对象重复出现在不同子空间上。

(3)特征选择方法

特征选择就是从特征集 $T=\{t_1,t_2,\cdots,t_s\}$ 中选择一个真子集 $T'=\{t_1,t_2,\cdots,t_s\}$,满足 $s'\ll s$。其中,s 为原始特征集的大小,s'为选择后的特征集大小。特征选择不需变换,而是从维度中启发式地选取一部分维度,删除不相关或冗余的属性,目标是找出最小属性集,使得数据类的概率分布尽可能接近使用所有属性得到的原分布。这种方法避免了挖掘结果难以解释的问题,并且属性数目的减少使得模式更易于理解[75]。Chen 等[76]使用德国和澳大利亚的 UCI 数据库的信用评分数据比较了 4 种属性过滤方法:线性判别分析法、决策树法、粗糙集法和 F 评分法。以 SVM 为基分类器进行实验,结果表明属性过滤算法对于分类精度的提高是有效的,但不同数据集使用的过滤算法有所差异。刘颖等[77]提出基于粒子群协同优化算法的供应链金融信用风险评价模型,该模型利用二进制粒子群算法优选特征子集,并对支持向量机参数进行协同优化。Huang 等[78]将遗传算法与 BPN、GP 与 C4.5 特征子集获取算法进行比较,以 SVM 作为基分类器,结果表明,遗传算法所选的子集在信用评价中有较好的分类精度。Wang 等[79]提出一种基于过滤方法和多群体遗传算法(MPGA)的信用评分特征选择混合模型,该方法在第一阶段将包装方法的思想引入到 3 种滤波方法中,以获得 MPGA 初始种群设置的一些重要先验信息,第二阶段利用 MPGA 的全局优化和快速收敛的特点来寻找最优特征子集。

结束语　近年来,严格的信贷增量控制致使实体经济,特别是中小企业融资困境日趋加剧。供应链金融的实质是为了解决相对弱势的中小企业的融资问题。与传统融资模式相比,供应链融资包含的资金关系错综复杂,评估指标动态多样,半结构、非结构化的海量金融数据使得数据分析方法较为复杂。大数据以一种前所未有的方式,通过对海量数据进行分析,来获得有巨大价值的产品和服务,也包括深刻的洞见[80]。在复杂的非均衡、噪声形态和非线性形变的情况下,高效、准确地对信用风险进行有效评价,一直以来都是一项极具挑战性的任务。本文从供应链金融数据的分布特征入手,分析影响信用风险分类模型的主要因素,大致包括信用数据的非均衡性与不对称性、信用数据的噪声和离群点的存在以及信用数据的非线性多维特征。非均衡性与不对称性问题的主要解决策略包括利用重采样、机器学习等方法来平衡样本。对于噪声离群点,其解决方法集中于聚类算法、距离离群点探测和密度离群点探测等方法。多维特征的降维方法主要包括线性降维、投影变换和特征选择方法。本文对供应链金融数据的分布特征进行深入探索,梳理信息获取的主要影响因素,挖掘信用评估数据中蕴含的内在信息与研究价值,为构建性能更为优越的信用评估模型奠定研究基础。下一步工作将在信用数据分布特征的基础上融合多种改进策略,构建适用于供应链金融领域的信用风险度量模型。

参 考 文 献

[1] TRUONG N,LI Z,VIRGINIA S,et al. Big data analytics in supply chain management: A state-of-the-art literature review [J]. Computers and Operations Research,2018,98:254-264.

[2] RICHARD L V,MATTHEW E,CARL W O. Big data: What it is and why you should care[M]. IDC Go-to-Market Services, 2011.

[3] RICHARD A T,PETRI T H. Big data applications in operations/supply-chain management: A literature review[J]. Computers & Industrial Engineering,2016,101:528-543.

[4] GANTZ J,REINSEL D. Extracting value from chaos[M]. IDC Go-to-Market Services,2011:1-12.

[5] HUANG Y Y,HANDFIELD R B. Measuring the benefits of ERP on supply management maturity model: a 'big data' method[J]. International Journal of Operation & Production Management,2015,35 (1):2-25.

[6] CHEN C L P,ZHANG C Y. Data-intensive applications, challenges, techniques and technologies: a survey on Big Data[J]. Information Sciences,2014,275(11):314-347.

[7] BABICEANU R F,SEKER R. Big Data and virtualization for manufacturing cyber-physical systems: a survey of the current status and future outlook[J]. Computers in industry,2016,81: 128-137.

[8] CHAO L M,XING C X,ZHANG Y. Data Science Studies: State-of-the-art and Trends[J]. Computer Science,2018,45(1): 1-13. (in Chinese)
朝乐门,邢春晓,张勇. 数据科学研究的现状与趋势[J]. 计算机科学,2018,45(1):1-13.

[9] XU H L,TANG S,MAO R,et al. Various Pivots Based Outlier Dectection Algorithm in Metric Space[J]. Chinese Journal of Computers,2017,40(12):2839-2855. (in Chinese)
许红龙,唐颂,毛睿,等. 基于多种支撑点的度量空间离群检测算法[J]. 计算机学报,2017,40(12):2839-2855.

[10] 袁荃. 基于供应链金融的中小企业融资决策研究[D]. 武汉:武汉大学,2010.

[11] Demica Limited Company. Research report: A study on the growth of supply chain finance, as evidenced by SCF[EB/OL]. http://www.demica.com.

[12] XIAO J,XUE S T,HUANG J,et al. A Semi-Supervised Co-Training Model for Customer Credit Scoring[J]. Chinese Journal of Management Science,2016,24(6):124-131. (in Chinese)
肖进,薛书田,黄静,等. 客户信用评估半监督协同训练模型研究[J]. 中国管理科学,2016,24(6):124-131.

[13] YANG J,ZHOU Y G. Credit risk spillovers among financial institutions around the global credit crisis: Firm-level evidence[J]. Management Science,2013,59(10):2343-2359.

[14] CHEN H,CHIANG R H,STOREY V C. Business intelligence and analytics: From big data to big impact[J]. MIS Quarterly, 2012,36(4):1165-1188.

[15] ARCHENAA J,ANITA E M. A survey of big data analytics in

healthcare and government[J]. Procedia Computer Science, 2015,50:408-413.

[16] VATRAPU R,MUKKAMALA R R,HUSSAIN A,et al. Social set analysis:A set theoretical approach to big data analytics[J]. IEEE Access,2016,4:2542-2571.

[17] KHAN Z,ANJUM A,SOOMRO K,et al. Towards cloud based big data analytics for smart future cities[J]. Journal of Cloud Computing,2015,4(1):2.

[18] FIOSINA J,FIOSINS M,MULLER J P. Big data processing and mining for next generation intelligent transportation systems [J]. Journal Teknologi,2013,63(3):21-38.

[19] SLEDGIANOWSKI D,GOMAA M,TAN C. Toward integration of Big Data,technology and information systems competencies into the accounting curriculum[J]. Journal of Accounting Education,2017,38:81-93.

[20] CERCHIELLO P,GIUDICI P. Big data analysis for financial risk management[J]. Journal of Big Data,2016,3(1):1-12.

[21] ZHAO N,ZHANG X F,ZHANG L J. Overview of Imbalanced Data Classification[J]. Chinese Journal of Computers,2018, 45(S1):22-27. (in Chinese)
赵楠,张小芳,张利军. 不平衡数据分类研究综述[J]. 计算机科学,2018,45(S1):22-27.

[22] DEBASHREE D,SAROJ K B,BISWAJIT P. Redundancy-driven modified Tomek-link based undersampling: A solution to class imbalance[J]. Pattern Recognition Letters,2017,93(1):3-12.

[23] YANG Z,ABHISHEK K S,KWOK L T. Imbalanced classification by learning hidden data structure[J]. IIE Transations, 2016,48(7):614-628.

[24] YI B H,ZHU J J,LI J. Imbalanced Data Classification on Micro-Credit Company Customer Credit Risk Assessment Using Improved SMOTE Support Vector Machine[J]. Chinese Journal of Mangement Science,2016,24(3):24-30. (in Chinese)
衣柏衡,朱建军,李杰. 基于改进 SMOTE 的小额贷款公司客户信用风险非均衡 SVM 分类[J]. 中国管理科学,2016,24(3):24-30.

[25] PIERRI F,STANGHELLINI E,BISTONI N. Risk analysis and retrospective unbalanced data[J]. Revstat-statistical Journal, 2016,14(2):157-169.

[26] LI S,SONG W F,QIN H,et al. Deep variance network : An iterative,improved CNN framework for unbalanced training datasets[J]. Pattern Recognition,2018,81:294-308.

[27] XIONG B Y,WANG G Y,DENG W B. Under-Sampling Method Based on Sample Weight for Imbalanced Data[J]. Journal of Computer Research and Development,2016,53(11):2613-2622. (in Chinese)
熊冰妍,王国胤,邓维斌. 基于样本权重的不平衡数据欠抽样方法[J]. 计算机研究与发展,2016,53(11):2613-2622.

[28] CHICLANA F,MATA F,PEREZ L G,et al. Type-1 OWA Unbalanced Fuzzy Linguistic Aggregation Methodology: Application to Eurobonds Credit Risk Evaluation[J]. International Journal of Intelligent Systems,2018,33(5):1071-1088.

[29] VAPNIK. The nature of statistical learning theory [M]. New York:Springer,1995:1-14.

[30] SHAO Y H,CHEN W J,ZHANG J J,et al. An efficient weighted Lagrangian twin support vector machine for imbalanced data classification [J]. Pattern Recognition,2014,47(9):3158-3167.

[31] CHENG Y Q. Credit Rating of Small Enterprises Based on Unbalanced Data [J]. Operations Research and Management Science,2016,25(6):181-189. (in Chinese)
程砚秋. 基于不平衡数据的小企业信用风险评价[J]. 运筹与管理,2016,25(6):181-189.

[32] GOMEZ C L,CAMPS V G,BRUZZONE L. Mean map kernel methods for semisupervised cloud classification[J]. IEEE Transactions on Geoscience and Remote Sensing,2010,48(1):207-220.

[33] XIA Z G,XIA S X,CAI S Y,et al. Semi-supervised Gaussian process classification algorithm addressing the class imbalance [J]. Journal on Communications,2013,34(5):42-51. (in Chinese)
夏战国,夏士雄,蔡世玉,等. 类不均衡的半监督高斯过程分类算法[J]. 通信学报,2013,34(5):42-51.

[34] LI X F,LI J,DONG Y F,et al. A New Learning Algorithm for Imbalanced Data-PCBoost[J]. Chinese Journal of Computers, 2012,35(2):202-209. (in Chinese)
李雄飞,李军,董元方,等. 一种新的不平衡数据学习算法 PCBoost[J]. 计算机学报,2012,35(2):202-209.

[35] LI K W,YANG L,LIU W Y,et al. Classification Method of Imbalanced Data Based on RSBoost[J]. Computer Science,2015, 42(9):249-252. (in Chinese)
李克文,杨磊,刘文英,等. 基于 RSBoost 算法的不平衡数据分类方法[J]. 计算机科学,2015,42(9):249-252.

[36] ZHU B,HE C Z,LI H Y. Research on Credit Scoring Model Based on Transfer Learning[J]. Operations Research and Management Science,2015,24(2):201-207. (in Chinese)
朱兵,贺昌政,李慧媛. 基于迁移学习的客户信用评估模型研究[J]. 运筹与管理,2015,24(2):201-207.

[37] CHANG Y C,CHANG K H,CHU H H,et al. Establishing decision tree-based short-term default credit risk assessment models[J]. Communications in Statistics-theory and Methods, 2016,45(23):6803-6815.

[38] SUN J,LEE Y C,LI H,et al. Combining B&B-based hybrid feature selection and the imbalance-oriented multiple-classifier ensemble for imbalanced credit risk assessment[J]. Technological and Economic Development of Economy,2015,21(3):351-378.

[39] LIU F,MAO Z Z,LI L. Outlier detection for control process data based on fuzzy ARHMM[J]. Chinese Journal of Scientific Instrument,2010,31(5):984-990. (in Chinese)
刘芳,毛志忠,李磊. 基于模糊自回归隐马尔可夫模型的控制过程异常数据检测[J]. 仪器仪表学报,2010,31(5):984-990.

[40] GRACES H,SBARBARO D. Outliers detection in environmental monitoring databases[J]. Engineering Application of Artifi-

cial Intelligence,2011,24(2):341-349.

[41] JIA R D,LIU J H,MAO Z Z,et al. Outlier detection for batch processes based on robust M-estimation[J]. Chinese Journal of Scientific Instrument,2013,34(8):1726-1731. (in Chinese)
贾润达,刘俊豪,毛志忠,等. 基于鲁棒 M 估计的间歇过程离群点检测[J]. 仪器仪表学报,2013,34(8):1726-1731.

[42] JIANG Z,ZHAN Y Z. Noise control and related algorithm for semi-supervised classification[J]. Journal of Jiangsu University (Natural Science Edition),2015,36(4):435-438. (in Chinese)
姜震,詹永照. 半监督分类中的噪声控制及相关算法[J]. 江苏大学学报(自然科学版),2015,36(4):435-438.

[43] WU J H,ZHANG Y,WANG X J. The Measurement Study of Corporate Bond Default Risk under the Information Disclosure Distortion[J]. Jouranl of Applied Statistics and Management,2017,36(1):175-190. (in Chinese)
吴建华,张颖,王新军. 信息披露扭曲下企业债券违约风险量化研究[J]. 数理统计与管理,2017,36(1):175-190.

[44] JIANG M F,TSENG S S,SU C M. Two-phase clustering process for outliers detection[J]. Pattern Recognition Letters,2001,22(6-7):691-700.

[45] ZHUANG H,ZHANG J,BROVA G,et al. Mining query-based subnetwork outliers in heterogeneous information networks[C]// IEEE International Conference on Data Mining,Piscataway. NJ:IEEE,2014:1127-1132.

[46] ZHU L,QIU Y Y,YU S,et al. A Fast KNN-Based MST Outlier Detection Method Chinese[J]. Journal of Computers,2017,40(12):2856-2870. (in Chinese)
朱利,邱媛媛,于帅,等. 一种基于快速 k-近邻的最小生成树离群检测方法[J]. 计算机学报,2017,40(12):2856-2870.

[47] PENG T,YANG N Y,XU Y B,et al. An Outlier Detection Method Based on Ranking and Clustering in Bi-typed Heterogeneous Network[J]. Acta Electronica Sinica,2018,46(2):281-288. (in Chinese)
彭涛,杨妮亚,徐原博,等. 双类型异质网中基于排序和聚类的离群点检测方法[J]. 电子学报,2018,46(2):281-288.

[48] LIU Y,WANG L M,JIANG J H,et al. SVM Credit Risk Evaluation Method Based on Eliminating Outliers[J]. Journal of Jilin University (Science Edition),2016,54(6):1395-1400. (in Chinese)
刘颖,王丽敏,姜建华,等. 基于离群点剔除的 SVM 信用风险评价方法[J]. 吉林大学学报(理学版),2016,54(6):1395-1400.

[49] KNORR E M,NG R T. Algorithms for mining distance-based outliers in large datasets[C]// Proceedings of the 24th International Conference on Very Large Data Bases. New York,USA,1998:392-403.

[50] WANG Y,PARTHASARATHY S,TATIKONDA S. Locality sensitive outlier detection:A ranking driven approach[C]// Proceedings of the IEEE 27th International Conference on Data Engineering. Hannover,Germany,2011:410-421.

[51] PILLUTLA M R,RAVAL N,BANSAL P,et al. LSH based outlier detection and its application in distributed setting[C]// Proceedings of the 20th ACM International Conference on Information and Knowledge Management. Glasgow,UK,2011:2289-2292.

[52] WANG X T,SHEN D R,BAI M,et al. BOD:An Efficient Algorithm for Distributed Outlier Detection[J]. Chinese Journal of Computers,2016,39(1):36-50. (in Chinese)
王习特,申德荣,白梅,等. BOD:一种高效的分布式离群点检测算法[J]. 计算机学报,2016,39(1):36-50.

[53] JIANG F,SUI Y F,CAO C G. Distance metrics and outlier detection in rough sets[J]. Control and Decision,2013,28(1):188-192. (in Chinese)
江峰,眭跃飞,曹存根. 粗糙集中的距离度量与离群点检测[J]. 控制与决策,2013,28(1):188-192.

[54] YAO X,YU L A. A fuzzy proximal support vector machine model and its application to credit risk analysis[J]. Systems Engineering-Theory & Practice,2012,32(3):549-554. (in Chinese)
姚潇,余乐安. 模糊近似支持向量机模型及其在信用风险评估中的应用[J]. 系统工程理论与实践,2012,32(3):549-554.

[55] LIU J L,LI J P,XU W X,et al. A Robust Weighted Adaptive LpLS-SVM Method for Credit Risk Assessment[J]. Chinese Journal of Management Science,2010,18(5):28-33. (in Chinese)
刘京礼,李建平,徐伟宣,等. 信用评估中的鲁棒赋权自适应 Lp 最小二乘支持向量机方法[J]. 中国管理科学,2010,18(5):28-33.

[56] BHADURI K,MATTHEWS B L,GIANNELLA C R. Algorithms for speeding up distance-based outlier detection[C]// Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Diego,USA,2011:859-867.

[57] BREUNIG M M. LOF:Identifying density-based local outliers [J]. ACM Sigmod Record,2015,29(2):93-104.

[58] JIN W,TUNG A K H,HAN J,et al. Ranking outliers using symmetric neighborhood relationship[J]. Lecture Notes in Computer Science,2006,3918:577-593.

[59] ZHOU S B,XU W X. Deviation-based local outlier detection algorithm[J]. Chinese Journal of Scientific Instrument,2014,35(10):2293-2298. (in Chinese)
周世波,徐维祥. 一种基于偏离的局部离群点检测算法[J]. 仪器仪表学报,2014,35(10):2293-2298.

[60] LIU Z T,XU J P,WU M,et al. Review of Emotional Feature Extraction and Dimension Reduction Method for Speech Emotion Recognition[J/OL]. Chinese Journal of Computers,http://kns.cnki.net/kcms/detail/11.1826.TP.20170813.1200.006.html. (in Chinese)
刘振焘,徐建平,吴敏,等. 语音情感特征提取及其降维方法综述[J/OL]. 计算机学报,http://kns.cnki.net/kcms/detail/11.1826.TP.20170813.1200.006.html.

[61] MENG D Y,XU C,XU Z B. A New Manifold Reconstruction Method Based on Isomap[J]. Chinese Journal of Computers,

2010,33(3):545-554. (in Chinese)
孟德宇,徐晨,徐宗本. 基于 Isomap 的流形结构重建方法[J]. 计算机学报,2010,33(3):545-554.

[62] ZHANG R C,DU Y B,XUE L G,et al. A hybrid large sample credit evaluation model based on combining similar samples[J]. Journal of Management Sciences in China,2018,21(7):77-90. (in Chinese)
张润驰,杜亚斌,薛立国,等. 基于相似样本归并的大样本混合信用评估模型[J]. 管理科学学报,2018,21(7):77-90.

[63] CHEN W S,DU Y K. Using Neural Networks and Data Mining Techniques for the Financial Distress Prediction Model[J]. Expert Systems with Applications,2009,36:4075-4086.

[64] PAN H P,ZHANG C Z. FEPA-An Adaptive Integrated Prediction Model of Financial Time Series[J]. Chinese Journal of Management Science,2018,26(6):26-38. (in Chinese)
潘和平,张承钊. FEPA-金融时间序列自适应组合预测模型[J]. 中国管理科学,2018,26(6):26-38.

[65] WEST D. Neural network credit scoring models[J]. Computer & Operations Research,2000,27:1131-1152.

[66] HUA Z,WANG Z,XU X,et al. Predicting Corporate Financial Distress Based on Integration of Support Vector Machine and Logistic Regression[J]. Expert Systems with Applications,2007,33(2):434-440.

[67] XIONG Z B. Research on Credit Evaluation Model Based on Nonlinear Principal Component Analysis[J]. The Journal of Quantitative & Technical Economics,2013(10):138-151. (in Chinese)
熊志斌. 基于非线性主成分分析的信用评估模型研究[J]. 数量经济技术经济研究,2013(10):138-151.

[68] ZHANG H X,MAO Z Z. Research of multidimensional time series credit evaluation based on gray-fuzz analysis model[J]. Journal of Management Sciences in China,2011,14(1):28-37. (in Chinese)
张洪祥,毛志忠. 基于多维时间序列的灰色模糊信用评价研究[J]. 管理科学学报,2011,14(1):28-37.

[69] ZHANG J,ZHANG B B. The Application of Generalized Semiparametric Additive Credit Score Model Based on Group-LASSO Method[J]. Journal of Applied Statistics and Management,2016,35(3):517-524. (in Chinese)
张娟,张贝贝. 基于 Group-LASSO 方法的广义半参数可加信用风险评分模型应用研究[J]. 数理统计与管理,2016,35(3):517-524.

[70] TENENBAUM J B,SILVA V,LANGFORD J C. A global geometric framework for nonlinear dimensionality reduction[J]. Science,2000,290(5500):2319-2323.

[71] LI F Y,DENG X. The Application Analysis of SVM Model Based on Isomap in the Credit Risk Assessment of Listed Companies[J]. Journal of Hebei University (Philosophy and Social Science),2013,38(1):102-107. (in Chinese)
李菲雅,邓翔. 等距特征映射的支持向量机模型在上市公司信用风险评估中的应用[J]. 河北大学学报(哲学社会科学版),2013,38(1):102-107.

[72] LIN F,YEH C C,LEE M Y. The use of hybrid manifold learning and support vector machines in the prediction of business failure[J]. Knowledge-Based Systems,2011,24(1):95-101.

[73] RIBEIRO B,VIEIRA A,DUARTE J,et al. Learning manifolds for bankruptcy analysis[M]// Advances in Neuro-Information Processing—ICONIP 2008. Berlin:Springer,2008:723-730.

[74] TONG G G,LI S W. Construction and Application Research of Isomap-RVM Credit Assessment Model[J]. Mathematical Problems in Engineering,2015,2015:1-7.

[75] XUE A R,YAO L,JU S G,et al. Survey of Outlier Mining[J]. Computer Science,2008,35(11):13-18. (in Chinese)
薛安荣,姚林,鞠时光,等. 离群点挖掘方法综述[J]. 计算机科学,2008,35(11):13-18.

[76] CHEN F L,LI F C. Combination of feature selection approaches with svm in credit scoring[J]. Expert System Application,2010,37:4902-4909.

[77] LIU Y,ZHANG L J,HAN Y N,et al. Credit Risk Evaluation Model of Supply Chain Finance Based on Particle Swarm Cooperative Optimization Algorithm[J]. Journal of Jilin University (Science Edition),2018,56(1):119-125. (in Chinese)
刘颖,张丽娟,韩亚男,等. 基于粒子群协同优化算法的供应链金融信用风险评价模型[J]. 吉林大学学报(理学版),2018,56(1):119-125.

[78] HUANG C L,CHEN M C,WANG C J. Credit scoring with a data mining approach based on support vector machines[J]. Expert System Application,2007,33:847-856.

[79] WANG D,ZHANG Z Q,BAI R Q,et al. A hybrid system with filter approach and multiple population genetic algorithm for feature selection in credit scoring[J]. Journal of Computational and Applied Mathematics,2018,329:307-321.

[80] HAGSTROM M. High-performance analytics fuels innovation and inclusive growth:Use big data,hyper connectivity and speed to intelligence to get true value in the digital economy[J]. Journal of Advanced Analytics,2012,2:3-4.