基于大数据最优特征子集的小企业信用风险预警研究

迟国泰，章彤
(大连理工大学 经济管理学院，辽宁 大连 116024)

目录

基于大数据最优特征子集的小企业信用风险预警研究

迟国泰，章彤

(大连理工大学 经济管理学院，辽宁 大连 116024)

摘要：信用评级是对客户的信用状况进行评价，信用评级结果不合理，会误导银行并产生信用风险。本研究的问题一是遴选出兼顾违约判别能力最强和指标冗余最小的指标组合，二是发掘出影响客户信用的关键状态。**本文创新与特色一是以**所有指标组合与违约状态之间的平均“互信息”最大为目标来保证遴选的指标组合具有最大的违约鉴别能力，**以**所有指标组合中的指标与指标之间的平均“互信息”最小为目标来保证在被遴选的指标组合中，指标间的信息冗余最小；**通过**这种指标组合最小冗余最大相关(mRMR)为目标建立了 0-1 规划模型、**遴选**出兼顾违约判别能力最大和冗余度最小的最优指标组合。**通过** 0-1 规划的全局最优解，**改变**了现有研究采用增量搜索算法仅仅会得到局部最优解的弊端。实证研究表明，劳动力人数、家庭人数/劳动力人数、居住年限、贷款人的家庭日常生活支出等 5 个指标构成的组合最优。**二是通过** χ^2 统计量把最优指标组合中的每一个数值型指标划分成不同区间或状态，并用虚拟变量编码表示每一个数值型和类别型指标的状态，以所有指标的不同的状态(而不是对指标数值)为自变量，以客户的真实的违约状态为因变量进行逻辑回归，构建基于指标状态赋权的信用评分方程，**并通过**逻辑回归方程的系数 β_i 大小揭示影响违约鉴别能力的信用指标关键状态，完善了现有研究只遴选指标组合，忽略揭示指标状态对违约鉴别能力影响的问题。对中国农户贷款的实证研究表明，当家庭人数与劳动力人数的比值在 3.5 到 4 之间时，农户贷款更容易违约；当农户历史上申请贷款次数少于 4 次时，农户贷款违约的人数更多；居住年限在 1 到 32 年之间的农户信用水平相对于其他居住年限的农户更好，而居住少于 1 年的农户信用水平相对更差。

关键词：信用评级；指标组合；最优指标组合；指标状态；互信息；0-1 规划

中图分类号：F830.56；O221.2　　**文献标识码：**A

1 引言

信用评级是金融机构对企业或个人的信用情况进行评估，进而为贷款决策和利率定价提供依据。信用评级须基于信用指标体系对客户违约状态进行判别，指标体系决定了信用评级模型的合理性，因而指标体系的遴选至关重要。

在信用指标遴选过程中，一种较广泛使用的方法是逐个选择违约鉴别能力强的指标，再将违约鉴别能力最强的前n个指标选出来作为指标组合[1]。这种方法固然可行，然而并不能保证选出的指标组合最优。由于指标与指标之间存在着相关性，单个指标的违约鉴别能力强，并不能保证指标组合的违约鉴别能力强，对于信用评级而言，指标组合的违约鉴别能力无疑是更重要的。如何遴选出一组具有最大违约判别能力的指标组合，显然更有意义也更具有挑战。

(1)信用评级中指标遴选的研究现状：

信用评级中指标遴选的方法主要分为 3 大类：单指标遴选、指标群遴选以及混合方法[1]。

单指标遴选方法主要基于单个指标的违约判别能力以及指标的相关性进行指标筛选，通过保留违约鉴别能力强的指标，删除违约鉴别能力弱的指标，达到精简指标组合的目的[1]。代表性的研究有：

在信用评级领域，很多学者将多种单指遴选方法用于指标遴选，并对比每种方法选出的指标在分类模型中的精度。Piramuthu(1999)提出了一种基于模糊的指标遴选方法，他将该方法与随机选择指标、Patrick-Fisher 概率性距离指标选择方法以及全指标的数据集进行对比，发现这种新的方法筛选出的指标在决策树分类模型下，比其他三种方法具有更好的分类精度[2]。Liu 和 Schumann(2005)将 ReliefF 方法、标准化信息增益方法、一致性方法三种单指标遴选方法以及基于分类模型的指标群遴选方法分别用于信用数据的指标遴选，并进行了对比，结果表明经过指标遴选的数据集在计算速度和判别精度上都有显著提高[3]。其他学者 Aryuni 和 Madyatmadja(2015)、Koutanaei 等(2015)也对多种应用于信用评级的单指标遴选方法进行了对比[4,5]。

还有一些学者将多种单指标遴选结合起来使用，综合每种方法的优点，寻求更好的指标组合。Chen(2012)分别采用卡方平方、ReliefF 算法、信息增益率以及信息增益等四种单指标遴选方法，并综合四种方法对指标的进行重要性排序[6]。Bouaguel 等(2013)、Sadatrasoul 等(2015)以及 Dahiya 等(2016)均提出了将多种单指标遴选方法综合使用的方法[7-9]。

现有研究的单指标遴选方法[2-9]不论是只考虑一种指标遴选方法还是综合考虑多种指标遴选方法，其弊端是仅考虑了单个指标对违约的判别能力，却没有考虑指标组合整体的违约鉴别能力，这样可能导致选出的指标组合整体违约鉴别能力不强。

指标群遴选方法是综合考虑指标组合的违约

基金项目：国家自然科学基金重点项目(71731003，71431002)；国家自然科学基金面上项目(71471027，71171031，71873103)；国家社科基金一般项目(16BTJ017)；国家自然科学基金青年科学基金项目(71601041)；辽宁省社科规划基金项目(L16BJY016)；辽宁经济社会发展重点课题(2015lslktzdian-05)；大连银行小企业信用风险评级系统与贷款定价项目(2012-01)；中国邮政储蓄银行总行小额贷款信用风险评价与贷款定价资助项目(2009-07)。

作者简介：迟国泰(1955-)，男，黑龙江海伦人，教授，博士生导师，博士，chigt@dlut.edu.cn，研究方向：信用评级；章彤(1990-)，男，管理科学与工程专业博士研究生，zhangtong19900227@126.com，研究方向：信用评级。

通讯作者：迟国泰

鉴别能力，而不考虑单个指标的特性，一次性挑选出一组具有最优鉴别能力的指标群[1]。代表性的研究有：

大部分指标群遴选方法是将指标组合的枚举算法与分类判别模型结合使用，选择出使得分类精度最高的指标组合。Huang 等(2010)基于支持向量机分类模型，提出了一种嵌入式递归指标遴选方法，并且在多分类的信用数据中采用，研究表明在加入这种嵌入式递归指标遴选方法后，多分类支持向量机比传统的多分类模型具有更好的精度[10]。Wang 等(2010，2012)先后将粗糙集方法与散点搜索以及 tabu 搜索两种算法结合，寻找整体条件熵最大的指标组合[11,12]。另外，Wang 和 Huang(2009)、Hajek 和 Michalak(2013)的研究基于信用评级数据，对比了多种指标群遴选方法与多种单指标遴选方法选择的指标组合对于违约的判别精度，研究表明指标群遴选方法比单指标遴选方法速度更慢但违约判别准确率更高[13-14]。

还有一些研究考虑了指标组合内的相关性，根据指标组合整体的信息含量最大进行指标群遴选。Wang 等(2017)基于社区发现提出了一种特征选择方法，根据指标之间的相关性构建带权网络，并通过社区发现算法找出网络中每个社区中信息量最大的指标构成信用指标组合，用于信用卡违约预测[15]。

一些研究将信用成本指标获取的成本、信用评分的收益等都考虑进模型。Maldonado 等(2017)提出了两种基于支持向量机的规划模型进行指标群遴选，并同时考虑了指标获取的成本[16]。Maldonado (2017)在另一篇文章中将信用评分的收益也纳入了模型中，构建了同时考虑成本和收益的指标群遴选模型[17]。

现有的研究的指标群遴选方法大部分都需要依赖一种枚举搜索算法和分类算法相结合，寻找判别精度最高的指标组合，由于很多枚举搜索算法由于算法的局限性只能找到局部最优解，很难找到真正最优的指标组合。

混合方法是先通过单指标遴选方法初步筛选指标，再通过指标群遴选方法进一步选出判别能力强的指标群，这种方法既能保证筛选速度也能保证指标的违约判别能力[1]。代表性的研究如下：

Chen 和 Li(2010)基于 UCI 公开数据集中的德国和澳大利亚信用数据，对比了线性判别分析、粗糙集、决策树、*F*-score 四类单指标遴选方法与基于支持向量机的指标群遴选方法进行结合，结果表明混合指标筛选方法具有更强的稳定性和更高的效率[18]。Waad 等(2013)提出了一种三阶段的信用评级指标遴选方法，第一阶段采用多种单指标遴选方法获得与违约相关性高的指标，第二阶段采用二次优化方法删除冗余的指标，第三阶段用基于指标群遴选方法获得判别能力最强的指标组合，研究表明在多分类模型中，他们的方法相对于其他单指标或者指标群遴选方法都具有更好的分类准确率[19]。另外，Oreski S 和 Oreski G(2014)、Stąpor 等(2016)、Wang 等(2017)均研究了不同的混合型指标遴选方法，并与多种单指标和指标群遴选方法进行了对比[20-22]。

现有研究的混合指标遴选方法[18-22]是在单指标遴选的基础上再进行指标群遴选，忽略了一些可能的指标组合，因此选出的指标组合也不能保证最优。

(2)信用评级中评价方法的研究现状

一是基于统计学的信用评价方法。Altman(1968)最早基于线性判别分析方法提出的著名的 Z-score 模型，该模型是 5 个财务指标的线性加权[23]。Grablowsky 和 Talley(1981)最先将 Probit 回归方法用于信用评级模型的赋权[24]。之后的研究者 Boyes 等(1989)、Duca 和 Whitesell(1995)均用 probit 回归构建了信用评价模型，对信用卡客户进行信用评估[25-26]。Hand 和 Henley(1997)、Dong 等(2010)分别通过了不同的逻辑回归方法构建了个人信用贷款的信用评价模型[27-28]。

二是基于机器学习的信用评价方法。West(2000)对比了五种神经网络方法用于信用评价，进而构建信用评价模型，并将神经网络方法与判别分析方法、逻辑回归方法、K-近邻方法、核密度估计方法、决策树方法的实证结果进行对比，发现径向基函数神经网络方法具有最高的判别精度[29]。Tsai 和 Wu(2008)采用了多层感知器神经网络方法构建信用评价模型[30]。Ong 等(2005)用遗传规划算法构建了信用评分模型，研究表明遗传规划算法构建的模型在预测方面优于人工神经网络模型、粗糙集方法、决策树方法、逻辑回归方法[31]。Abdou(2009)同样采用了遗传规划算法，发现遗传规划算法相对于 Probit 分析、证据权重方法等其他方法具有更高的平均正确分类率(ACC)以及更低的错判成本(EMC)[32]。

上述信用评价方法[23-32]均是研究信用评价指标与客户违约状态之间的关系，却很少有考虑信用评价指标的各种状态(例如，年龄作为一个信用评价指标，而各年龄段就是指标对应的各种状态)与违约的影响。

Ding 和 Peng(2005)提出了基于指标组合信息最小冗余最大相关(Minimum redundancy maximum relevance，mRMR)标准，用于遴选基因序列的指标群，他们采用一种简单的增量搜索算法，寻找最小冗余最大相关的指标组合[33]。然而由于这种增量搜索的方法的局限性，最终找到的指标组合并非全局最优。

本文借鉴 mRMR 标准，以指标组合的信息冗余度最小及违约鉴别能力最大为目标，以指标是否选入为决策变量，构建了 0-1 规划模型进行指标群遴选，通过 $\chi 2$ 统计量把最优指标组合中的每一个数值型指标划分成不同区间或状态，并用逻辑回归对各指标经过虚拟变量编码的状态变量进行拟合，

构建了以指标状态为自变量的信用评价模型。

2 信用评级体系构建的原理

2.1 信用评级体系构建的难点及突破难点的思路

本文涉及的问题有二：

一是对于 n 个指标，总共能构成 2^n-1 个指标组合，当 n 较大，例如 n=40 时，共有 1.1×10^{12} 个指标组合(2^{40}-1≈1.1×10^{12})，如何从众多的指标组合中遴选出满足最小冗余最大相关(mRMR)标准的指标组合，属于数学中的 N-P 难问题。

二是揭示对于任何一个指标的不同状态，哪一个状态对违约判别具有关键影响。

解决第一个问题的思路是以所有指标组合与违约状态之间的平均“互信息”最大为目标来保证遴选的指标组合具有最大的违约鉴别能力，以所有指标组合中的指标与指标之间的平均“互信息”最小为目标来保证在被遴选的指标组合中，指标间的信息冗余最小；通过这种指标组合最小冗余最大相关(mRMR)为目标建立了 0-1 规划模型、遴选出兼顾违约判别能力最大和冗余度最小的最优指标组合。

解决第二个问题的思路是通过 χ^2 统计量把最优指标组合中的每一个数值型指标划分成不同区间或状态，并用虚拟变量编码表示每一个数值型和类别型指标的状态，以所有指标的不同的状态(而不是对指标数值)为自变量，以客户的真实的违约状态为因变量进行逻辑回归，通过逻辑回归方程的系数 β_i 揭示指标组合中不同状态对违约鉴别能力的影响。

2.2 指标群遴选的两个命题及其证明思路

命题 1：由单个违约鉴别能力强的指标构成的指标组合，违约鉴别能力不一定强。

证明思路：从所有指标中选出与本模型指标数量相同的单个违约鉴别能力最大的指标构成的指标组合，计算指标组合中违约相关度与冗余度之差，并与本模型所选指标进行对比，若相对本模型更差，说明由单个违约鉴别能力强的指标构成的指标组合，违约鉴别能力不一定强。

命题 2：指标组合的个数不是越多越好。

证明思路：依次控制指标组合的个数，分别选择出指标个数从 1 到 n 的最优的 n 个指标组合，对比这 n 个指标组合相关度与冗余度的比值。若随着指标个数增多，指标组合的相关度与冗余度的比值反而有减小的趋势，则说明指标组合的个数不是越多越好。

3 基于 mRMR 最优指标组合遴选的信用评级模型构建

3.1 指标数据的处理

3.1.1 数值型数据区间划分方法

指标数据有数值型数据(例如收入、人口数等)和类别型数据(例如职业、学历等)两种类型。

为了测算指标 X 与违约状态 Y 之间的互信息 I(X,Y)，则需要把数值型指标转换成类别型指标。由此，将所有指标都变成类别型指标，就可以通过互信息参数进行指标组合的遴选。

数值型指标转化的思路，是根据相邻两个区间中违约客户与非违约客户的人数构建卡方统计量 χ^2，χ^2 越小说明相邻两个区间中违约与非违约客户的比例越相似，则应该合并相邻两个区间；重复合并χ^2 最小的相邻区间，直到满足区间划分的停止标准。

区间划分停止的标准为：所有相邻两个区间的 χ^2 均通过显著性水平为 α 的χ^2 检验且区间的数量达到最大值 Q_{max} 或最小值 $Q_{\min}$。通常χ^2 检验的显著性水平 α 取值为 99%，最大值 Q_{max} 取 4，最小值 $Q_{\min}$ 取 2[34]。

步骤：Step1 将指标数据按从小到大顺序排列，将数值相同的数据划分为同一个区间，例如下文表 4 第 3 列的上边数值小，下边数值大。

Step2：构建相邻区间的χ^2

设：E_{ij}-相邻两个区间中第 i 个区间(i=1 或 i=2)中第 j 类客户预期人数，本文只有两类客户，当 j=1 时，代表非违约客户，当 j=2 时代表违约客户。N_{ij}-相邻区间的第 i 个区间(i=1 或 i=2)中第 j 类(j=1 或 j=2)客户人数。

$$E_{ij}=\frac{\sum_{i=1}^{2}N_{ij}}{\sum_{i=1}^{2}\sum_{j=1}^{2}N_{ij}}\times\sum_{j=1}^{2}N_{ij} \tag{1}$$

式(1)等式右端第一项的分子 $\sum_{i=1}^{2}N_{ij}$ 是相邻两个区间(第 1 个区间和第 2 个区间)中第 j 类客户人数。

式(1)等式右端第一项的分母 $\sum_{i=1}^{2}\sum_{j=1}^{2}N_{ij}$ 是相邻两个区间中两类客户的人数之和。

式(1)等式右端第一项 $\sum_{i=1}^{2}N_{ij}/\sum_{i=1}^{2}\sum_{j=1}^{2}N_{ij}$ 是相邻两个区间中第 j 类客户人数与两类客户人数之和的商，表示相邻两个区间中第 j 类客户的比率。

式(1)是相邻两个区间中第 j 类客户的比率与第 i 个区间中人数的乘积，表示第 i 个区间中第 j 类客户预期人数。

设：χ^2-相邻两个区间的卡方统计量。其他变量的含义同式(1)。

$$\chi^2=\sum_{i=1}^{2}\sum_{j=1}^{2}\frac{(N_{ij}-E_{ij})^2}{E_{ij}} \tag{2}$$

式(2)等号右边连加项中分子$(N_{ij}-E_{ij})^2$ 是第 i 个区间段中第 j 类客户人数与第 i 个区间段中第 j 类客户预期人数之差的平方，表示第 i 个区间内第 j 类客户的实际人数与预期人数的差距，$(N_{ij}-E_{ij})^2$ 越大说明两个区间内违约客户与非违约客户的比例差距较大。

式(2)等号右边服从χ^2 分布，是卡方统计量[11]。

式(2)连加号中的分子$(N_{ij}-E_{ij})^2$越大，则式(2) χ^2数值越大，即相邻两个区间段中违约与非违约客户人数分布的差别越大，则相邻区间越应该分开。

Step3：合并区间

根据式(1)和式(2)，计算所有两个相邻区间的χ^2，将其中χ^2最小的两个相邻区间进行合并。

Step4：区间合并停止

重复 Step3，区间划分的数量在[2,4]之间且所有相邻两个区间的χ^2通过显著水平为 99%χ^2检验，即χ^2大于 6.64 则区间合并停止[34]。

需要指出的是，由于每个区间对应指标数据的一个状态，当区间划分个数过多时，指标数据的状态过多。由于实证样本数量限制，将实证样本代入下文的逻辑回归中会出现指标状态总数过多，回归无法收敛的情况，因此为了控制区间划分的个数，本文设定的最大区间划分个数为 Q_{max}=4 个。

Step5：对区间进行标记

按区间的排列顺序，给第一个区间标记为 1，给第二个区间标记为 2，依此类推。

数值型数据通过以上区间划分方法可以转化为类别型数据。

本文通过区间划分方法根据指标数据与违约状态的关联关系对指标数据进行区间划分，以每个区间作为一个类别，将数值型数据转化为类别型数据，无需事先确定指标数据与客户信用状况的相关性，且保证数据能够代入下文的公式(3)中进行计算。

3.1.2 类别型指标分类标记

对类别型指标进行类别标记，以便之后代入模型进行处理。表 1 是类别型指标的状态标识规则，表 1 第 1 列是序号，第 2 列是类别型指标，第 3 列是类别型指标各个状态，第 4 列是类别标识。

表 1 类别型指标类别标识

Tab.1 The classification mark of qualitative indicators

(1)序号	(2)指标	(3)指标状态	(4)类别标记
1	学历	1.本科及以上	1
		2.大专	2
		3.高中	3
		4.初中	4
		5.小学	5
		6.其他	6
…		…	…
14	贷款用途	1.种植业、养殖业等生产费用贷款	1
		…	…
		4.其他	4

3.2 基于 mRMR 标准的指标群遴选整数规划模型

3.2.1 指标组合与违约状态的相关性 $D(X_i,Y)$

设 $I(X_i,Y)$-第 i 个指标 X_i 和客户违约状态 Y 之间的互信息。u-第 i 个指标的类别标识。以下文表 3 中第 9 行第 9 个指标供养人口 X_9 为例，经过上文 3.1.1 区间划分后，供养人口数被分为 2 个区间，此时 u 有 1 和 2 两个取值，当供养人口数为 0 至 4 人时，u=1。供养人口数为 5 人及以上时，u=2。v-客户的违约状态，违约时记为 1，非违约时记为 0。$p(u,v)$-客户第 i 个指标类别标识为 u，违约状态为 v 时的概率。$p(u)$-客户第 i 个指标类别标识为 u 的概率。$p(v)$-客户违约状态取值为 v 的概率。

则客户第 i 个指标 X_i 和违约状态 Y 之间的互信息如式(3)所示[33,35]。

$$\boldsymbol{I}(\boldsymbol{X}_i,\boldsymbol{Y})=\sum_{\boldsymbol{u}\in\boldsymbol{X}_i,\boldsymbol{v}\in\boldsymbol{Y}}\boldsymbol{p}(\boldsymbol{u},\boldsymbol{v})\log_2\frac{\boldsymbol{p}(\boldsymbol{u},\boldsymbol{v})}{\boldsymbol{p}(\boldsymbol{u})\boldsymbol{p}(\boldsymbol{v})} \tag{3}$$

式(3)连加号里面的第一项 $p(u,v)$是客户第 i 个指标取值 u 的同时对应的违约状态为 v 的概率。

式(3)连加号里面的第二项对数右边的分母 $p(u)p(v)$是客户第 i 个指标取值 u 的概率与客户违约状态为 v 的概率之积。

根据概率论的知识，当指标 X_i 和违约状况 Y 不相关时，有 $p(u,v)=p(u)p(v)$。当两者相关性大，则 $p(u,v)$就相比于 $p(u)p(v)$越大，即 $p(u,v)$与 $p(u)p(v)$的比值就越大，式(3)的数值就越大，因此可以用式(3)指标 X_i 与违约状态 Y 之间的互信息值来表示它们之间的相关性。

设：$D(S,Y)$-指标组合 S 与违约状态 Y 的相关性。S-指标组合。m-指标组合 S 中指标的个数。$I(X_i,Y)$-第 i 个指标 X_i 和客户违约状态 Y 之间的互信息。则指标组合 S 与违约状态之间的相关性如式(4)所示[33,35]。

$$\boldsymbol{D}(\boldsymbol{S},\boldsymbol{Y})=\frac{1}{\boldsymbol{m}}\sum_{\boldsymbol{X}_i\in\boldsymbol{S}}\boldsymbol{I}(\boldsymbol{X}_i,\boldsymbol{Y}) \tag{4}$$

式(4)是指标组合 S 中所有指标与违约状态 Y 之间互信息的平均值，根据式(3)，单个指标的互信息 $I(X_i,Y)$能够表示指标 X_i 与违约状态 Y 之间的相关性，因此指标组合中所有指标与违约状态之间互信息的平均值能表示指标组合 S 与违约状态 Y 之间的相关性。

3.2.2 指标组合的冗余度 $R(S)$

设 $I(X_i,X_j)$-第 i 个指标 X_i 和第 j 个指标 X_j 之间的互信息。u-第 i 个指标的类别标识(此类别标识与式(3)中类别标识的含义相同)。v-第 j 个指标的类别标识。$p(u,v)$-客户第 i 个指标类别为 u，第 j 个指标类别为 v 时的概率。$p(u)$-客户第 i 个指标类别标识为 u 的概率。$p(v)$-客户第 j 个指标类别标识为 v 的概率。则客户第 i 个指标和第 j 个指标之间的互信息如式(5)所示[33,35]。

$$\boldsymbol{I}(\boldsymbol{X}_i,\boldsymbol{X}_j)=\sum_{\boldsymbol{u}\in\boldsymbol{X}_i,\boldsymbol{v}\in\boldsymbol{X}_j}\boldsymbol{p}(\boldsymbol{u},\boldsymbol{v})\log_2\frac{\boldsymbol{p}(\boldsymbol{u},\boldsymbol{v})}{\boldsymbol{p}(\boldsymbol{u})\boldsymbol{p}(\boldsymbol{v})} \tag{5}$$

式(5)与式(3)相似，式(3)是指标 X_i 和客户违约状态 Y 之间的互信息，表示指标 X_i 和客户违约状态 Y 之间的相关关系，而式(5)是指标 X_i 和指标 X_j 之间的互信息，表示指标之间的相关关系。

设 $R(S)$-指标组合 S 的冗余度。S-指标组合。m-指标组合 S 中的指标个数。$I(X_i,X_j)$-第 i 个指标 X_i 和第 i 个指标 X_j 之间的互信息。则指标组合 S 的冗

余度如式(6)所示[33]。

$$\boldsymbol{R}(\boldsymbol{S})=\frac{1}{\boldsymbol{m}^2}\sum_{\boldsymbol{X}_i,\boldsymbol{X}_j\in\boldsymbol{S}}\boldsymbol{I}(\boldsymbol{X}_i,\boldsymbol{X}_j) \tag{6}$$

式(6)是指标组合S中所有指标与指标之间互信息的平均值，根据式(5)，每一对指标之间的互信息$I(X_i,X_j)$能够表示指标X_i与指标X_j之间的相关性，而两个指标之间相关越大说明两个指标越能够相互替代，即两个指标之间存在冗余，因此指标组合中所有指标与指标之间互信息的平均值能够表示指标组合S的冗余度。当指标组合冗余度较大，说明指标之间存在相关性，这种情况下需要删除部分指标，减少冗余度。

3.2.3 基于 mRMR 标准的指标群遴选整数规划模型构建

(1)目标函数的建立

设$\Phi(S_0,Y,\boldsymbol{c})$-以$S_0$为初始指标群，并加入决策变量$\boldsymbol{c}$之后的指标组合最小冗余最大相关目标函数。$S_0$-初始指标群，即尚未经过指标遴选的原始指标群；$\boldsymbol{c}$-指标群 S_0 中指标是否被选入的决策向量，$\boldsymbol{c}=(c_1,\ldots c_i\ldots,c_n)$，$\boldsymbol{c}^{\mathrm{T}}$是$\boldsymbol{c}$的转置，$c_i$表示指标$i$是否被选入指标体系，当指标$c_i$被选入体系时，$c_i=0$，反之$c_i=1$。$I^{XY}$-所有$n$个指标与违约状态之间互信息构成的行向量，如式(8)所示，向量中第i个元素由上文式(3)计算。n-初始指标群S_0中指标的总数。I^{XX}-所有指标与指标之间互信息构成的$n\times n$矩阵，矩阵中第i行第j列的元素是第i个指标与第j个指标之间的互信息，如式(9)所示。$I(X_i,Y)$-第i个指标与违约状态之间的互信息。$I(X_i,X_j)$-第i个指标和第j个指标之间的互信息。则最小冗余最大相关指标群遴选的目标函数可以写成式(7)的形式。

$$\boldsymbol{Max}\ \Phi(\boldsymbol{S}_0,\boldsymbol{Y},\boldsymbol{c})=\frac{\boldsymbol{I}^{XY}\mathbf{c}^{\boldsymbol{T}}}{\sum_{i=1}^{n}\boldsymbol{c}_i}\Big/\frac{\mathbf{c}\boldsymbol{I}^{XX}\mathbf{c}^{\boldsymbol{T}}}{(\sum_{i=1}^{n}\boldsymbol{c}_i)^2} \tag{7}$$

$$\boldsymbol{I}^{XY}=\big(\boldsymbol{I}(\boldsymbol{X}_1,\boldsymbol{Y}),\boldsymbol{I}(\boldsymbol{X}_2,\boldsymbol{Y}),...,\boldsymbol{I}(\boldsymbol{X}_n,\boldsymbol{Y})\big) \tag{8}$$

$$\boldsymbol{I}^{XX}=\begin{pmatrix}\boldsymbol{I}(\boldsymbol{X}_1,\boldsymbol{X}_1) & \mathrm{L} & \boldsymbol{I}(\boldsymbol{X}_1,\boldsymbol{X}_n)\\ \mathrm{M} & \mathrm{O} & \mathrm{M}\\ \boldsymbol{I}(\boldsymbol{X}_n,\boldsymbol{X}_1) & \mathrm{L} & \boldsymbol{I}(\boldsymbol{X}_n,\boldsymbol{X}_n)\end{pmatrix} \tag{9}$$

式(7)等式右边被除数的分子$\boldsymbol{I}^{XY}\boldsymbol{c}^{\mathrm{T}}$是所有指标与违约状态之间互信息$I(X_i,Y)$和决策变量$c_i$乘积的总和。当指标$i$被选时，$c_i=1$，则指标的互信息参与求和，反之，当指标$i$未被选时，$c_i=0$，指标的互信息与 0 相乘后数值为 0，即不参与求和。因此，式(7)等式右边第一项分子能够表示被选指标组合互信息的总和，与式(4)中分子意义相同。

式(7)右边被除数的分母$\sum_{i=1}^{n}\boldsymbol{c}_i$是决策变量$c_i$数值的总和，当指标$i$被选入时，$c_i$取 1，反之取 0，因此所有$c_i$的加和即为被选入的指标组合中指标数目，相当于式(4)中分母m。

式(7)等号右边被除数是被选指标组合与违约状态之间互信息总和与指标组合中指标数目的商，与式(4)含义相同，因此能够表示被选指标组合与违约状态的相关性，是指标组合相关性$D(S,Y)$增加了决策变量c_i之后的形式。

式(7)等式右边除数的分子$c\boldsymbol{I}^{XX}c^{\mathrm{T}}$是所有指标$X_i$与指标$X_j$之间的互信息$I(X_i,X_j)$与决策变量$c_i$以及$c_j$的乘积之和，当指标$i$和指标$j$均被选入，即$c_i$和$c_j$均取值 1 的时候，指标$i$和指标$j$之间的互信息才会参与求和，否则，当指标$i$和指标$j$中有一个未被选入时，它们之间的互信息均不会参与求和。因此，式(7)等式右边第二项的分子能够表示被选指标与指标之间互信息的总和，与式(6)中分子意义相同。

式(7)等式右边除数的分母$(\sum_{i=1}^{n}\boldsymbol{c}_i)^2$是决策变量$c_i$数值总和的平方，相当于式(6)中分母$m^2$。

式(7)等式右边除数是被选指标组合中指标与指标之间互信息总和与指标组合中指标数目平方的商，与式(6)含义相同，因此能够表示被选指标组合的冗余度，是指标组合冗余度$R(S)$增加了决策变量c_i之后的形式。

式(7)表示被选指标组合违约相关性与冗余度的比值最大的目标函数，保证了指标体系与违约的相关性最大且冗余度最小，其中向量$\boldsymbol{c}$是决策向量，其中每一个元素c_i对应着指标是否被选入指标组合，因此，总存在一组$\boldsymbol{c}$能够使得目标函数取值最大。

(2)约束条件的建立

$$1\le\sum_{i=1}^{n}\boldsymbol{c}_i\le\boldsymbol{n} \tag{10}$$

c_i大于等于 1 表示至少应该有一个指标被选入指标体系；c_i小于等于n表示指标体系中指标的数目不大于总数n。

以所有指标组合与违约状态之间的平均“互信息”最大为目标来保证遴选的指标组合具有最大的违约鉴别能力，以所有指标组合中的指标与指标之间的平均“互信息”最小为目标保证遴选的指标组合信息冗余最小，通过以指标组合的最小冗余最大相关(mRMR)为目标，即式(7)，以指标是否被选入作为决策变量，建立 0-1 规划模型，遴选出兼顾违约判别能力最大和冗余度最小的最优指标组合。通过 0-1 规划求解全局最优解，改变了现有研究采用增量搜索算法仅仅会得到局部最优解的弊端。

3.3 信用评分模型的构建

信用评分模型用来计算客户的信用评分。

首先对类别型指标进行虚拟变量编码，对类别型指标进行虚拟变量编码是为了将指标状态拆分开，这样在回归时可以对指标的不同状态进行回归，能够揭示指标的不同状态对信用评分的影响。

对数据进行虚拟变量编码后，将编码后的指标

数据代入逻辑回归模型中拟合，得出每一个客户的违约概率，进而基于违约概率计算客户信用得分。

3.3.1 虚拟变量编码

虚拟变量编码是将具有 m 个类别的指标拆分成 m 个指标的过程[25]。

表 2 指标虚拟变量编码

Tab.2 The dummy encoding of classification variables

(1)序号	(2)农户编号	(3)性别	(4)分类标识 X_4	(5)虚拟变量编码	
				(5-1)X_{4_1}	(5-2)X_{4_2}
1	农户 1	男	1	1	0
2	农户 2	女	2	0	1
…	…	…	…	…	…
2043	农户 2043	男	1	1	0
2044	农户 2044	男	1	1	0

以第 4 个指标性别 X_4 为例，如表 2 所示进行虚拟变量编码的说明。表 2 第 1 列是序号，第 2 列是农户编号，第 3 列是农户性别，第 4 列是指标类别标识，第 5 列是虚拟变量编码后的数值。

以农户 1 为例，农户 1 的性别为男，其分类标识为 1，则在第 1 行(5-1)列填上 1，在(5-2)列填上 0。

对于农户 2，其性别为女，分类标识为 2，则在第 2 行(5-1)列填上 0，在(5-2)列填上 1。

以此类推，对 X_4 性别进行虚拟变量编码，即可将 X_4 拆分为两个变量 X_{4_1} 和 X_{4_2}，若 X_4=1，则 (X_{4_1},X_{4_2})=(1,0)。若 X_4=2，则(X_{4_1},X_{4_2})=(0,1)。

有多个类别标识的指标通过同样方式进行虚拟变量编码。例如第 2 个指标学历 X_2 有 6 个状态，则可将 X_2 拆分为$(X_{2_1},X_{2_2},X_{2_3},X_{2_4},X_{2_5},X_{2_6})$，当 X_2 为“初中”，标识为 4 时，虚拟变量编码为(0,0,0,1,0,0)。

任何具有 m 个类别的变量均能用虚拟变量进行编码，进而将一个变量拆分为 m 个取值为 0 或 1 的变量。

3.3.2 逻辑回归信用评分模型

将通过上文虚拟变量编码处理之后的变量代入逻辑回归模型与违约状态进行拟合，获得客户违约的概率 P[36]。用 1-P 计算出非违约的概率，并将信用得分设置为 100(1-P)。

设 $P^{(j)}$-客户 j 的违约概率。β_{i_ki}-第 i 个指标第 k_i 个状态通过逻辑回归之后对应的系数。$x^{(j)}_{i\_ki}$-第 j 个客户第 i 个指标第 k_i 个状态的取值。$Score^{(j)}$-第 j 个客户的信用得分。

$$P^{(j)}=1/(1+exp(-((\beta_{1\_1}x^{(j)}_{1\_1}+\beta_{1\_2}x^{(j)}_{1\_2}+\ldots+\beta_{1\_k1}x^{(j)}_{1\_k1})+\ldots+(\beta_{i\_1}x^{(j)}_{i\_1}+\beta_{i\_1}x^{(j)}_{i\_1}+\ldots+\beta_{i\_ki}x^{(j)}_{i\_ki})+\ldots+(\beta_{n\_1}x^{(j)}_{n\_1}+\beta_{n\_2}x^{(j)}_{n\_2}+\ldots+\beta_{n\_kn}x^{(j)}_{n\_kn}))) \quad (11)$$

式(11)与现有研究的区别是大部分现有研究是将指标数据与客户的违约状态进行回归，而式(11)是将指标虚拟化编码后的数据与违约状态进行逻辑回归，从而反映指标的不同状态与客户违约状态之间的关系。

用 100 乘以非违约概率，获得式(12)。

$$Score^{(j)}=100(1-P^{(j)}) \quad (12)$$

将式(11)代入式(12)整理得式(13)，即为最后的信用评分模型：

$$Score^{(j)}=100/(1+exp(((\beta_{1\_1}x^{(j)}_{1\_1}+\beta_{1\_2}x^{(j)}_{1\_2}+\ldots+\beta_{1\_k1}x^{(j)}_{1\_k1})+\ldots+(\beta_{i\_1}x^{(j)}_{i\_1}+\beta_{i\_2}x^{(j)}_{i\_2}+\ldots+\beta_{i\_ki}x^{(j)}_{i\_ki})+\ldots+(\beta_{n\_1}x^{(j)}_{n\_1}+\beta_{n\_2}x^{(j)}_{n\_2}+\ldots+\beta_{n\_kn}x^{(j)}_{n\_kn})))) \quad (13)$$

用虚拟变量编码表示每一个数值型和类别型指标的状态，以所有指标的不同的状态(而不是对指标数值)为自变量，以客户的真实的违约状态为因变量进行逻辑回归，通过逻辑回归方程的系数 β_{i_ki} 揭示指标组合中不同状态对违约鉴别能力的影响，完善了现有研究只遴选指标组合，忽略揭示指标状态对违约鉴别能力影响的问题。

3.4 信用等级的划分

思路：(1)初步划分为 9 个信用等级。将客户按照式(13)计算的信用得分降序排列，并初步划分为 9 个等级。

(2)调整并确定最终信用等级。每个信用等级的损失率为等级内所有客户的总应收未收本息占总应收本息的比重。因此**通过**改变每个信用等级对应的信用得分上下限，可以改变等级内的客户数，进而改变信用等级客户群的总应收本息、总应收未收本息，引起损失率的改变。**通过对**每个债信等级的得分上下限不断调整，**直至**得到满足“信用等级越高，损失率越低”标准的信用等级划分结果。

上述信用等级划分思路是我们团队被授权的**中华人民共和国发明专利**[37][38]的思路，计算过程由计算机完成。

4.实证分析

4.1 样本选取

本文选取了某国有商业银行在中国 28 个省市自治区分支行的的农户贷款数据[4]，共 2 044 笔贷款，每笔贷款的数据包括年龄、学历、年净收入、居民消费价格指数等 44 个指标以及违约状态、应收本息、应收未收本息，如表 3 所示。

表 3 第 a 列是序号，第 b 列是准则层，第 c 列是指标名称，第 d 列是指标类型，在 44 个指标中，有 14 个指标是类别型指标，其他 30 个指标为数值型指标，除了 44 个指标以外，表 3 第 44 至 46 行分别给出了农户的违约状态、应收本息以及应收未收本息的数据。

表 3 第 1 至 2 044 列是农户借据的原始数据，其中前 1 816 列为非违约借据，第 1 816-2 044 列为违约借据。

表 3 第 2 045 至 4 088 列为所有农户数据转化为类别型之后的数据，均用类别标识表示，数据转化过程将在下文 4.2 中介绍。

表 3 第 e 列的指标是否被选择的 0-1 变量标识，0 表示指标不被选择，1 表示指标被选入指标体系，计算方法将在下文 4.3 中介绍。

4.2 指标数据的预处理

本文中对指标数据进行数值型数据分类转化

的预处理有两个目的：一是进行数值型数据分类转化后，将数据由连续型变成离散型，在下文 4.3 中才便于计算互信息值。二是数据分类转化之后，每一个数值表示指标的一个状态，再经过虚拟变量编码后代入下文 4.4 中的逻辑回归方程中时，能够回归出每个指标状态的相关系数，进而反映指标状态对违约的影响程度。

表 3 农户信用评级原始指标数据与类别转化后数据

Tab.3 The raw data and classified data of farmer

(*a*)序号	(*b*)准则层	(*c*)指标	(*d*)指标类型	2 044 笔借据的指标原始数据 u_{ij}					2 044 笔借据的指标分类标识数据 x_{ij}					(e)0-1 变量 c_i
				1 816 笔非违约借据			228 笔违约借据		1 816 笔非违约借据			228 笔违约借据		
				(1)农户 1	…	(1 816) 农户 1 816	(1 817)农户 1 817	…	(2 044)农户 2 044	(2 045)农户 1	…	(3 860) 农户 3 860	(3 861)农户 3 861	…	(4 088)农户 4 088	
1	C_1 基本情况	X_1 年龄	数值	57	…	36	45	…	24	2	…	1	1	…	1	0
2		X_2 学历	类别	小学		初中	小学		初中	5		4	5		4	0
…		…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
9		X_9 供养人口	数值	0	…	2	4	…	2	1	…	1	1	…	1	0
10		X_{10} 家庭人数/劳动力人数	数值	1	…	2	3	…	4	1	…	1	1	…	3	1
11	C_2 还款能力	X_{11} 居住状况	类别	自有住房	…	自有住房	自有住房	…	自有住房	1	…	1	1	…	1	0
12		X_{12} 居住年限	数值	10	…	4	11	…	6	2	…	2	2	…	2	1
…		…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
29		X_{29} 银行存款	数值	0	…	0	0	…	0	1	…	1	1	…	1	0
30	C_3 还款意愿	X_{30} 民间借贷	类别	有	…	有	有	…	有	1	…	1	1	…	1	0
…		…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
32		X_{32} 历史上申请贷款次数	数值	0	…	1	0		0	1	…	1	1		1	1
…		…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
34		X_{34} 社会信誉	类别	无负面评价	…	无负面评价	无负面评价	…	无负面评价	4	…	4	4	…	4	0
35	C_4 联保保证	X_{35} 是否有保证	类别	否	…	是	否	…	是	2	…	1	2	…	1	0
…		…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
38		X_{38} 联保关系	类别	关系好	…	无联保	关系好	…	无联保	1	…	4	1	…	4	0
39	C_5 宏观环境	X_{39} 农村家庭人均纯收入	数值	3 502.9	…	9 257.9	3 502.9	…	7 356.5	1	…	3	1	…	3	0
…		…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
44		X_{44} 恩格尔系数	数值	0.373	…	0.364	0.373	…	0.379	1	…	1	1	…	1	0
45	——	**是否违约**	—	非违约	…	非违约	违约	…	违约	0	…	0	1	…	1	—
46	——	**应收本息**	—	0		0	58.115		9.154	—						
47	——	**应收未收本息**	—	53 910		52 358.75	10 573.75		10 530.7	—						

4.2.1 数值型数据区间划分

其中对于数值型数据，采用上文 3.1.1 中的区间划分算法对数值型数据进行区间划分，并转化为类别型数据。本文基于区间划分算法的参数设置为：显著性水平为 99%，最大区间划分个数 Q_{max}=4 个，最小区间划分个数为 Q_{min}=2 个[34]。

以第 9 个指标供养人口 X_9 为例，说明通过区间划分算法进行数据区间划分的计算过程。

表 4 供养人口 X_9 初始区间划分

Tab.4 Initial binning of X_9

(1)序号	(2)编号	(3)供养人口 X_9	(4)是否违约	(5)初始区间	(6)非违约数	(7)违约数
1	农户 1	0	0	1	405	56
…	…	…	…			
461	农户 2 043	0	1			
462	农户 3	1	0	2	926	121
…	…	…	…			
1 508	农户 2 042	1	1			
1 509	农户 2	2	0	3	369	43
…	…	…	…			
1 920	农户 2 044	2	1			
1 921	农户 4	3	0	4	99	4
…	…	…	…			
2 023	农户 2 026	3	1			
2 024	农户 45	4	0	5	15	2
…	…	…	…			
2 040	农户 1 919	4	1			
2 041	农户 52	5	0	6	2	2
…	…	…	…			
2 044	农户 1 968	5	1			

Step1：初始区间划分

将表 3 第 9 行 1-2 044 列中 X_9 供养人口的指标

数据按由小到大的顺序排列之后，列入表 4 第 3 列中。表 4 第 1 列是序号，第 2 列是对应的农户编号，第 4 列是农户对应的违约状态(0-非违约，1-违约)。

表 4 第 5 列为初始区间的序号，按照供养人口数相同划分为同一个区间。例如表 4 第 3 列第 1 至 461 行供养人口数均为 0，因此划为一个区间，在表 4 第 5 列第 1 行中并标上序号“1”;表 4 第 3 列第 462 至 1 508 行供养人口数均为 1，划为同一个区间，在表 4 第 5 列第 2 行中并标上序号“2”；依此类推，将供养人口划分为 6 个初始区间，如表 4 第 5 列。

表 4 第 6 列为每一个区间对应的非违约客户数。对于第 1 个区间，统计表 4 第 3 列第 1-461 行中非违约客户的人数，为 405 个，列入表 4 第 6 列第 1 行中；对于第 2 个区间，统计表 4 第 3 列 462 行至 1 508 行中非违约客户个数，为 926 个，列入表 4 第 6 列第 2 行中；依此类推，分别统计后 4 个区间中非违约客户人数，分别列入表 4 第 6 列 3 至 6 行。

表 4 第 7 列为每一个区间对应的违约客户数。对于第 1 个区间，统计表 4 第 3 列第 1-461 行中违约客户的人数，为 56 个，列入表 4 第 7 列第 1 行中；对于第 2 个区间，统计表 4 第 3 列 462-1 508 行中违约客户个数，为 121 个，列入表 4 第 7 列第 2 行中；依此类推，分别统计后 4 个区间中违约客户人数，分别列入表 4 第 7 列 3-6 行。

Step2：第 2 次区间划分

Step2.1：计算相邻两个区间中第 *i* 个区间(*i*=1 或 *i*=2)中第 *j* 类(*j*=1 为非违约客户，*j*=2 为违约客户)客户预期人数 E_{ij}。

以第 1 和第 2 相邻两个区间第 1 个区间非违约客户的预期人数 E_{11} 计算为例，涉及 4 个参数如下。

第 1 个参数是区间 1 中非违约客户的个数 N_{11}=405，来自表 4 第 1 行第 6 列；

第 2 个参数是区间 1 中违约客户的个数 N_{12}=56，来自表 4 第 1 行第 7 列；

第 3 个参数是区间 2 中非违约客户的个数 N_{21}=926，来自表 4 第 2 行第 6 列；

第 4 个参数是区间 2 中违约客户的个数 N_{22}=121，来自表 4 第 2 行第 7 列。

将上述 4 个参数代入式(1)中，则区间 1 中非违约客户的预期人数 E_{11} 为：

$E_{11}=(N_{11}+N_{21})(N_{11}+N_{12})/(N_{11}+N_{12}+N_{21}+N_{22})=(405+926)\times(405+56)/(405+56+926+121)=406.89$。

同理可以得到相邻两个区间中第 1 个区间中违约客户预期人数 E_{12}=54.11，相邻两个区间中第 2 个区间中非违约客户预期人数 E_{21}=924.11，相邻两个区间中第 2 个区间中违约客户预期人数 E_{22}=122.89。

表 5 供养人口 X_9 区间合并过程

Tab.5 the binning merge process of feeding population X_9

(1)供养人口 X_9	(2)初始区间划分及χ^2		(3)第 2 次区间划分及χ^2		(4)第 3 次区间划分及χ^2		(5)第 4 次区间划分及χ^2		(6)第 5 次区间划分及χ^2	
	(2-1)	(2-2)	(3-1)	(3-2)	(4-1)	(4-2)	(5-1)	(5-2)	(6-1)	(6-2)
0	1		1		1		1		1	
1	2	$\chi^2(1,2)^{(1)}$=**0.108**								
2	3	$\chi^2(2,3)^{(1)}$=0.372	2	$\chi^2(1,2)^{(2)}$=**0.539**						
3	4	$\chi^2(3,4)^{(1)}$=4.267	3	$\chi^2(2,3)^{(2)}$=4.267	2	$\chi^2(1,2)^{(3)}$=5.696		$\chi^2(1,2)^{(4)}$=**4.782**		
4	5	$\chi^2(4,5)^{(1)}$=1.908	4	$\chi^2(3,4)^{(2)}$=1.908	3	$\chi^2(2,3)^{(3)}$=**1.908**	2			
5	6	$\chi^2(5,6)^{(1)}$=3.07	5	$\chi^2(4,5)^{(2)}$=3.07	4	$\chi^2(3,4)^{(3)}$=3.07	3	$\chi^2(2,3)^{(4)}$=12.988	2	$\chi^2(1,2)^{(5)}$=6.102

Step2.2：计算初始区间划分的χ2

以表 4 第 5 列中第 1 和第 2 两个相邻区间的 $\chi^2(1,2)^{(1)}$计算为例，进行说明，其中共涉及 8 个参数：

其中第 1-4 个参数为 Step2.1 中用到过的区间 1 中非违约客户的个数 N_{11}、区间 1 中违约客户的个数 N_{12}、区间 2 中非违约客户的个数 N_{21}、区间 2 中违约客户的个数 N_{22}。

另外第 5-8 个参数为 Step2.1 计算得的区间 1 中非违约客户的预期个数 E_{11}、区间 2 中违约客户的预期个数 E_{12}、区间 2 中违约客户的预期个数 E_{21}、区间 2 中违约客户的预期个数 E_{22}。

将上述 8 个参数的数值代入式(2)，可以计算出区间 1 与区间 2 的$\chi^2(1,2)^{(1)}$。

$\chi^2(1,2)^{(1)}=(N_{11}-E_{11})^2/E_{11}+(N_{12}-E_{12})^2/E_{12}+(N_{21}-E_{21})^2/E_{21}+(N_{22}-E_{22})^2/E_{22}=(405-406.89)^2/406.89+(56-54.11)^2/54.11+(926-924.11)^2/924.11+(121-122.89)^2/122.89=0.108$

其中$\chi^2(1,2)^{(1)}$代表区间 1 和区间 2 的卡方统计量，上标(1)代表第 1 次区间划分，下文类似，不再撰述。

将第 1 次区间划分后区间 1 和区间 2 的 $\chi^2(1,2)^{(1)}$=0.108 列入表 5 第(2-2)列第 1、2 行之间的单元格中，表示区间 1 和区间 2 的χ^2。

同理可以计算出第 1 次区间划分后区间 2 和区间 3 的$\chi^2(2,3)^{(1)}$=0.372，区间 3 和区间 4 的$\chi^2(3,4)^{(1)}$=4.26，区间 4 和区间 5 的$\chi^2(4,5)^{(1)}$=1.90，区间 5 和区间 6 的$\chi^2(5,6)^{(1)}$=3.07，分别填入表 5 第(2-2)列中相应的单元格中。

Step2.3：对χ^2最小的相邻区间合并

对表 5 第(2-2)列中第 1 次区间划分后相邻区间的χ^2大小进行对比，其中第 1 次区间划分后区间 1 和区间 2 的$\chi^2(1,2)^{(1)}$=0.108 数值最小，说明区间 1 和区间 2 中违约与非违约客户的比例最相近，因此对区间 1 和区间 2 进行合并。合并之后供养人口数为 0 和 1 被并为区间 1，在表 5 第(3-1)列第 1 行填入区间标识“1”。

其他区间不进行合并，只对表 5 第(2-1)列中第 1 次区间划分的标识依次减 1 作为第 2 次区间划分的标识，填入第(3-1)列 2 至 5 行。

Step3：最终区间划分结束

第 2 次区间划分是以初始区间划分结果，即表 5 第(2-1)列，为基准，计算相邻区间的χ^2，合并χ^2最小的相邻区间，结果在第(3-1)列。

第 3 次区间划分则以第 2 次区间划分的结果，即第(3-1)列为基准，计算相邻区间的χ^2，合并最小χ^2对应的相邻区间，区间划分新标识填入第(4-1)列，χ^2填入第(3-2)列。

以此类推，仿照 Step2 和 Step3 的方法，每次以上一次区间划分的结果为基准，计算相邻区间的χ^2，并将χ^2最小的相邻区间进行合并，将第 3、4、5 次区间划分计算的χ^2分别填入表 5 第(3-2)、(4-2)、(5-2)列，区间划分结果分别填入(5-1)、(6-1)列。

直到第 5 次区间划分结束后，只剩下 2 个区间，计算相邻区间的$\chi^2(1,2)^{(5)}$=6.102，填入表 5 中(6-2)列。虽然χ^2数值还未超过 99%显著水平的 6.64，但区间划分已经达到了的最小区间数 2，故不再继续合并区间，此时区间划分停止。

至此第 9 个指标供养人口 X_9 的区间划分结束。最终指标供养人口 X_9 被划分为[0,5)和[5,+∞]两个区间，分别填入表 6 中第(2-1)和(2-2)行第 3 列。区间 1 标识为“1”，区间 2 标识为“2”，分别填入第(2-1)和(2-2)行第 4 列。

类推，可以得到其他 30 个数值型指标数据的区间划分，将其区间划分结果填入表 6 其他行和列。

表 6 农户数值型数据类别型转化标准

Tab.6 The binning standard of numeric farmer data

(1)序号		(2)指标	(3)指标区间划分	(4)类别标识
1	(1-1)	X_1 年龄(岁)	[-∞,48)	1
	(1-2)		[48, +∞]	2
2	(2-1)	X_9 供养人口(人)	[0,5)	1
	(2-2)		[5,+∞]	2
3	(3-1)	X_{10} 家庭人数/劳动力人数	[-∞,3.5)	1
	(3-2)		[3.5,4)	2
	(3-3)		[4,-∞]	3
…		…	…	…
30	(30-1)	X_{44} 恩格尔系数	[-∞,0.399)	1
	(30-2)		[0.399,0.405)	2
	(30-3)		[0.405, +∞]	3

4.2.2 指标数据转化为类别标识

根据表 6 中数值型指标的区间划分对应的类别标识以及上文表 1 类别型指标的分类标识，将表 3 第 1-2 044 列前 44 行的数据转化为类别标识，填入第 2 045-4 088 列前 44 行。

以农户 1 的两个指标为例说明表 3 第 2 045-4 088 列的来龙去脉：

对于数值型数据，以“年龄”为例。在表 3 第 1 行第 1 列农户 1 的年龄 X_1 为 57 岁，在表 6 中第(1-2)行第 4 列找到对应的年龄[48,+∞]的区间，类别标记为 2，把“2”这个标记填入表 3 第 1 行第 2 045 列中。

表 3 中的其他数值型指标，根据指标的名称和数值在表 6 中找到对应的标识，填入表 3 第 2 045 至 4 088 列其他数值型指标对应的行。

对于类别型数据，以“学历”为例。对于表 3 第 2 行的指标“学历”，指标为类别型指标，根据表 1 第 1 行第 2 和 3 列中“学历”的类别对应的标识，将表 3 第 2 行 1-2 044 列的“学历”状态转化为类别标识，填入表 3 第 2 行 2 045 至 4 088 列。

表 3 中的其他类别型指标，仿照“学历”在表 1 中找到对应的标识，填入表 3 第 2 045 至 4 088 列其他类别型指标对应的行。

由此得到了表 3 前 44 行第 2 045 到 4 088 列的全部数据。

4.3 基于整数规划的最小冗余最大相关指标群遴选模型构建

4.3.1 指标与违约状态之间互信息 $I(X_i,Y)$的计算

根据表 3 中第 2 045 至 4 088 列的数据，通过式(3)，计算每一个指标 X_i 与违约状态 Y 之间的互信息值 $I(X_i,Y)$。

以第 9 个指标供养人口 X_9 与违约状态 Y 之间互信息值 $I(X_9,Y)$的计算为例，进行说明。

X_9 分类标记后的数据对应表 3 第 9 行第 2 045 至 4 088 列，共两种取值 1 和 2；而违约状态共 2 种取值 0 和 1。它们的组合共 2×2=4 种情况，分别为：(1,0)，(1,1)，(2,0)，(2,1)。用 $C_{1,0}$ 表示供养人口数类别标识为 1 且非违约的客户数，根据表 3 第 9 行和第 45 行 2 045 至 4 088 列的数据，可统计得 $C_{1,0}$=1 814。同理可以统计得供养人口数类别标识为 1 且违约的客户数 $C_{1,1}$=226，供养人口数类别标识为 2 且非违约的客户数 $C_{2,0}$=2，供养人口数类别标识为 2 且违约的客户数 $C_{2,1}$=2。

$I(X_9, Y)$的计算共涉及 8 个参数：

第 1 个参数为 $p(1,0)$，表示供养人口数类别标识为 1 且客户非违约的概率，由供养人口数类别标识为 1 且客户非违约的人数除以农户总数获得，即 $p(1,0)=C_{1,0}/2\ 044=1\ 814/2\ 044=0.887$。

同理可以得到参数 2 到参数 4：

供养人口数类别标识为 1 且客户违约的概率 $p(1,1)=C_{1,1}/2\ 044=226/2\ 044=0.111$。

供养人口数类别标识为 2 且客户非违约的概率 $p(2,0)= C_{2,0}/2\ 044=2/2\ 044=0.001$。

供养人口数类别标识为 2 且客户违约的概率 $p(2,1)=C_{2,1}/2\ 044=0.001$。

第 5 个参数 $p(X_9=1)$是供养人口数类别标识为 1 的概率，其为供养人口数类别标识为 1 时违约和非违约概率之和，即 $p(X_9=1)=p(1,0)+p(1,1)=$

0.887+0.111=0.998。

第 6 个参数 $p(X_9=2)$时供养人口数标识为 2 时得概率，其为供养人口数类别标识为 2 时违约与非违约概率之和，即 $p(X_9=2)=p(2,0)+p(2,1)=0.001+0.001=0.002$。

第 7 个参数 $p(Y=0)$是非违约客户的概率，即 $p(Y=0)=p(1,0)+p(2,0)=0.887+0.001=0.888$。

第 8 个参数 $p(Y=1)$是非违约客户的概率，即 $p(Y=1)=p(1,1)+p(2,1)=0.111+0.001=0.112$。

将以上 8 个参数代入中式(3)中，可以计算出供养人口 X_9 与违约状态之间的互信息值。

$I(X_9,Y)=p(1,0)\log_2(p(1,0)/p(X_9=1)p(Y=0))+p(1,1)\log_2(p(1,1)/p(X_9=1)p(Y=1))+p(2,0)\log_2(p(2,0)/p(X_9=2)p(Y=0))+p(2,1)\log_2(p(2,1)/p(X_9=2)p(Y=1))$

$=0.887\times\log_2(0.887/(0.998\times0.888))+\ldots+0.001\times\log_2(0.001/(0.002\times0.112))=0.001\ 3$

将 $I(X_9,Y)$的数值填入表 7 第 45 行第 9 列。

同理可以计算其他各指标与违约状态的互信息数值，计算过程通过 Matlab2018a 编程完成，计算结果列入表 7 第 45 行其他列。

把表 7 第 45 行的数据用 1×44 的向量 I^{XY} 表示，如式(14)所示。

$$I^{XY}=(0.000\ 04,\ldots,0.0013,\ldots,0.005\ 97) \quad (14)$$

向量 I^{XY} 的第 i 个元素代表第 i 个指标与违约状态之间的互信息值。

4.3.2 指标与指标之间互信息 $I(X_i,X_j)$的计算

上文 4.3.1 中计算的是指标供养人口 X_9 与违约状态 Y 之间的互信息。

这里计算的是指标 X_i 与指标 X_j 之间的互信息 $I(X_i,X_j)$。

他们都是两个要素之间的互信息计算，因此计算公式相同。故仿照(1)的计算过程，可以计算出指标与指标之间的互信息，填入表 7 前 44 行。

表 7 指标与违约状态之间互信息以及指标与指标之间互信息

Tab.7 The mutual information values between indicators and default status

序号	指标	指标					
		(1)X_1 年龄	…	(9)X_9 供养人口	…	(43)X_{43} 居民储蓄存款余额	(44)X_{44} 恩格尔系数
1	X_1 年龄	0.742 19	…	0.002 39	…	0.002 55	0.000 05
2	X_2 学历	0.018 62	…	0.011 31	…	0.034 34	0.021 84
3	X_3 婚姻状况	0.002 39	…	0.310 72	…	0.008 39	0.003 69
…	…	…	…	…	…	…	…
43	X_{43} 居民储蓄存款余额	0.002 55	…	0.008 39	…	0.82881	0.216 75
44	X_{44} 恩格尔系数	0.000 05	…	0.003 69	…	0.216 75	0.712 18
45	指标与违约状态的互信息 $I(X_i,Y)$	0.000 04	…	0.001 3	…	0.006 49	0.005 97

把表 7 前 44 行的数据形成 44×44 的矩阵 I^{XX}，如式(15)所示。

$$\boldsymbol{I}^{XX}=\begin{pmatrix}0.74219 & \text{L} & 0.00005\\ \text{M} & \text{O} & \text{M}\\ 0.00005 & \text{L} & 0.71218\end{pmatrix} \quad (15)$$

其中矩阵 I^{XX} 的第 i 行第 j 列的元素对应第 i 个指标与第 j 个指标之间的互信息值。

4.3.3 最小冗余最大相关的指标组合遴选目标函数构建

式(7)的计算共涉及 3 个参数的计算：

第 1 个参数 I^{XY} 是指标与违约状态之间的互信息矩阵，由上文式(14)的常数向量给出。

第 2 个参数 n 是原始指标的总个数，由表 3 第 c 列前 44 行可知共 44 个指标，n=44。

第 3 个参数 I^{XX} 是指标与指标之间的互信息矩阵，由上文式(15)的常数矩阵给出。

将上述三个参数代入式(7)可得式(16)。

$$\boldsymbol{Max}\ \Phi(\boldsymbol{S}_0,\boldsymbol{Y},\boldsymbol{c})=\frac{(0.00004,0.00283,\ldots,0.00597)\boldsymbol{c}^{T}}{\sum_{i=1}^{n}\boldsymbol{c}_i}\Big/\frac{\boldsymbol{c}\begin{pmatrix}0.74219 & \text{L} & 0.00005\\ \text{M} & \text{O} & \text{M}\\ 0.00005 & \text{L} & 0.71218\end{pmatrix}\boldsymbol{c}^{T}}{(\sum_{i=1}^{n}\boldsymbol{c}_i)^2} \quad (16)$$

4.3.4 约束条件的构建

本文采用中国农户数据中总指标为 44 个，因此将 n=44 代入式(10)可得式(17)。

$$1\le\sum_{i=1}^{44}\boldsymbol{c}_i\le 44，\ 且\ c_i=1\ 或\ c_i=0 \quad (17)$$

则式(17)为约束条件，其中 c_i=1 代表指标被选入指标体系，否则不被选入。

4.3.5 求解最优指标组合

求解以式(16)为目标函数，以式(17)为约束条件构成的 0-1 规划模型，得到由 0 和 1 构成的决策变量向量 $\boldsymbol{c}$，将向量 $\boldsymbol{c}$ 列入表 3 第 e 列。

在表 3 第 e 列中有 5 个指标对应的 c_i 值为 1，他们分别是：X_8 劳动力人数、X_{10} 家庭人数/劳动力人数、X_{12} 居住年限、X_{18} 贷款人的家庭日常生活支出、X_{32} 历史上申请贷款的次数。则这 5 个指标构成了最小冗余最大相关的最优指标组合。

由表 3 第 b 列和第 e 列可知决定农户的信用状况准则层为 C_1 基本状况、C_2 还款能力以及 C_3 还款意愿。

4.4 基于违约状态的信用评分方程的构建

4.4.1 指标数据虚拟变量编码

将最优指标组合的数据进行虚拟变量编码，填入表 8 前 14 行第 1 至 2 044 列。

表 8 第 a 列是序号。

第 b 列是准则层。

第 c 列的前 5 行是 5 个最优指标，后 2 行是客户违约状态和客户信用得分；其中客户违约状态来

自表 3 第 45 行 2 045 至 4 088 列。客户信用得分在下文 4.4.2 计算。

第 *d* 列是指标状态变量，上文 3.3.1 虚拟变量编码规则，将具有 *m* 个类别的指标变量拆分成了 *m* 个不同的变量。以指标家庭人数/劳动力人数 X_{10} 为例，查询表 6 第 3 行第 4 列，当指标状态标识为“1”，则状态变量记为 X_{10_1}，列入表 8 第 3 行第 *d* 列。当指标状态标识为“2”，则指标状态变量记为 X_{10_2}，列入表 8 第 4 行第 *d* 列。同理将指标 X_{10} 的其他状态变量和另外 4 个指标的状态变量均根据上述方式依次标记，列入表 8 第 *d* 列其他行。

第 *e* 列是指标状态对应的数值区间，是根据第 *d* 列指标状态变量对应的状态标识，在表 6 中查询得到相应指标状态的划分区间，填入表 8 第 *e* 列。

表 8 第 1-2 044 列是各指标状态变量进行虚拟变量编码之后的数值。以农户 1 为例进行说明，在表 3 第 10 行第 1 列查得其家庭人数/劳动力人数为 1，属于表 8 第 3 行第 *e* 列中家庭人数/劳动力人数的第 1 个指标状态区间[-∞,3.5)，因此在表 8 第 1 列状态变量 X_{10_1} 对应的第 3 行标记上 1，其他两个状态变量 X_{10_2} 和 X_{10_3} 对应的第 3 行和第 4 行标记 0。

同理将表 3 第 1-2 044 列中 5 个最优指标的其他数据均根据表 8 第 *e* 列的区间划分转变为虚拟变量编码，填入表 8 第 1-2 044 列 1-14 行其他位置。

表 8 第 *f* 列是通过逻辑回归模型拟合的结果，其中(*f*-1)列是系数名称、(*f*-2)列是系数值，具体计算将在下文 4.4.2 中介绍。

表 8 农户信用指标数据虚拟变量编码

Tab.8 The Dummy Encoding of farmer classified data

(*a*)序号	(*b*)准则层	(*c*)指标	(*d*)指标状态变量	(*e*)指标状态	2 044 笔借据的指标虚拟变量编码后数据						(*f*)逻辑回归结果	
					(1)农户 1	…	(1 954)农户 1 954	…	(2 044)农户 2 044		(*f*-1)系数	(*f*-2)系数值
1	C_1 基本情况	X_8 劳动力人数	$X_{8\text{-}1}$	[-∞,7)	1	…	1	…	1		$\beta_{8\text{-}1}$	-0.072
2			$X_{8\text{-}2}$	[7,-∞]	0		0		0		$\beta_{8\text{-}2}$	0.613
3		X_{10} 家庭人数/劳动力人数	$X_{10\text{-}1}$	[-∞,3.5)	1	…	1	…	0		$\beta_{10\text{-}1}$	-0.350
4			$X_{10\text{-}2}$	[3.5,4)	0		0		0		$\beta_{10\text{-}2}$	1.022
5			$X_{10\text{-}3}$	[4,-∞]	0		0		1		$\beta_{10\text{-}3}$	-0.130
6	C_2 还款能力	X_{12} 居住年限	$X_{12\text{-}1}$	[0,1)	0	…	0	…	0		$\beta_{12\text{-}1}$	1.428
7			$X_{12\text{-}2}$	[1,32)	1		0		1		$\beta_{12\text{-}2}$	-1.264
8			$X_{12\text{-}3}$	[32,33)	0	…	1	…	0		$\beta_{12\text{-}3}$	0.948
9			$X_{12\text{-}4}$	[33,-∞]	0		0		0		$\beta_{12\text{-}4}$	-0.571
10		X_{12} 贷款人的家庭日常生活支出	$X_{18\text{-}1}$	[-∞,800)	0	…	0	…	0		$\beta_{18\text{-}1}$	1.011
11			$X_{18\text{-}2}$	[800,2400)	0		0		0		$\beta_{18\text{-}2}$	0.602
12			$X_{18\text{-}3}$	[2400,-∞]	1		1		1		$\beta_{18\text{-}3}$	-1.072
13	C_3 还款意愿	X_{32} 历史上申请贷款次数	$X_{32\text{-}1}$	[-∞,4)	1	…	1	…	1		$\beta_{32\text{-}1}$	0.654
14			$X_{32\text{-}2}$	[4,-∞]	0		0		0		$\beta_{32\text{-}2}$	-0.113
15	——	**是否违约**	—	—	0	…	1	…	1		——	
16	——	**信用得分**	$Score^{(j)}$		89.14		47.30		89.14		——	

4.4.2 信用评分方程的构建

以表 8 中前 14 行 1 至 2 044 列的数据作为自变量数据，表 8 第 1 行 1 至 2 044 列的数据作为因变量数据，代入到逻辑回归模型中拟合，得出模型系数 β[16]，分别列入表 8 中第(*f*-2)列中。

将表 8 第(*f*-2)列前 14 行数据代入公式(13)，得到信用评分方程如式(18)所示。

$$Score^{(j)}=100/(1+\exp(((-0.072X_{8\_1}+0.613X_{8\_2})+(-0.35X_{10\_1}+1.022X_{10\_2}-0.13X_{10\_3})+(1.428X_{12\_1}-1.264X_{12\_2}+0.948X_{12\_3}-0.571X_{12\_4})+(1.011X_{18\_1}-0.602X_{18\_2}-1.072X_{18\_3})+(0.654X_{32\_1}-0.113X_{32\_2})))) \quad (18)$$

将表 8 第 1 至 14 行第 1 至 2 044 列数据代入式(18)可以分别计算出 2 044 个农户的信用评分，按顺序列入表 8 第 18 行第 1 至 2 044 列。

4.4.3 关键状态分析

由于方程(18)的回归系数在分母中，因此当回归系数大于 0 时，表示指标状态与信用评分是负相关，说明处于该指标状态的客户信用较好，反之，当回归系数小于 0 时，指标状态与信用评分为正相关，说明处于该指标状态的客户倾向于违约。

(1)在准则层 C_1 基本情况中，家庭人数与劳动力人数比值在 3.5 和 4 之间时，倾向于违约

由表 8 第 *f* 列第 4 行指标状态 X_{10_2} 的回归系数 $\beta_{10\text{-}2}$=1.022 可知，准则层 C_1 中指标“家庭人数/劳动力人数”数值处于 3.5 和 4 之间时，对信用评分的影响最大，是准则层 C_1 中最关键的信用状态。

根据式(18)，由于 $\beta_{10\text{-}2}$=1.022>0，说明家庭人数/劳动力人数处于 3.5 和 4 之间时，即状态变量 X_{10_2}=1 时，此时信用得分数值减少越多，说明这类客户倾向于违约。

(2)在准则层 C_2 还款能力中，居住年限在 1 年以内时，倾向于违约

由表 8 第 *f* 列第 6 行指标状态 X_{12_1} 的回归系数 $\beta_{12\text{-}1}$=1.428 可知，准则层 C_2 中指标“居住年限”数值处于 0 到 1 之间时，对信用评分的影响最大，是准则层 C_2 中最关键的信用状态。

根据式(18)，由于 $\beta_{12\text{-}2}$=1.428>0，说明农户的居住年限处于 0 到 1 之间时，即状态变量 X_{12_2}=1 时，此时信用得分数值减少越多，说明这类客户更容易违约。

(3)在准则层 C_3 还款意愿中，历史上申请贷款次数 4 次以下时，倾向于违约

由表 8 第 *f* 列第 13 行指标状态 X_{32_1} 的回归系数 $\beta_{32\text{-}1}$=0.654 可知，准则层 C_3 中指标“历史上申请贷款次数”在 4 次以下时，对信用评分的影响最大，是准则层 C_3 中最关键的信用状态。

根据式(18)，由于 $\beta_{32\text{-}2}$=0.654>0，说明农户的历史申请贷款次数少于 4 次时，即状态变量 X_{32_1}=1 时，此时信用得分数值减少越多，说明这类客户更容易违约。

(4)在所有状态中，居住年限处于 1 到 32 年时，信用状况较好

由表 8 第 *f* 列第 7 行指标状态 X_{12_1} 的回归系数 $\beta_{12\text{-}1}$=-1.246<0 可知，农户的居住年限在 1 至 32 年间时，对信用评分的增加最大，说明这类客户信用状况较好。

4.5 信用等级划分

本文通过我们团队获得的国家授权发明专利[37-38]进行信用等级的划分。

按照表 8 第 16 行的本模型计算的信用评分 $Score^{(j)}$从高到低，将贷款企业重新进行排序，并初步划分为 9 个等级，根据表 3 第 46 行的农户应收未收本息数据和第 47 行的应收本息数据分别算出每一个等级的损失率 *LR*[37-38]。通过调整等级得分上下限，每个等级对应的损失率 *LR* 随之改变，总会找到满足“信用等级越高、损失率越低”的最优划分结果。

划分结果如图 2 所示。图 2 中，每个等级对应的横轴长度表示这个等级对应的损失率大小。

由图 2 可知，本研究的信用等级划分结果**满足**“信用等级越高、损失率越低”的评级本质规律。

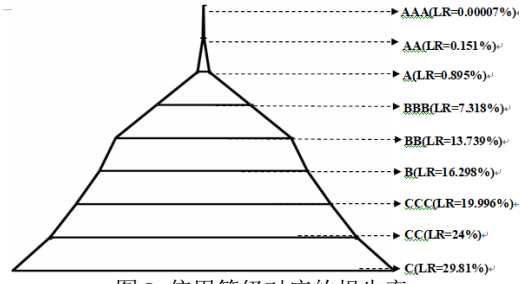

图 2 信用等级对应的损失率

Fig.2 *LR* corresponded to nine credit ratings

4.6 信用指标体系的对比分析

本文通过对 6 个指标体系进行对比，来证明本文指标体系的优越性，以及证明下文 4.7 中的两个命题。

(1)6 个指标体系的定义

指标体系 1: 本研究 5 个指标构成的指标体系。体系 1 中指标对应的指标，由上文表 3 第 *e* 列给出，将指标个数填入表 9 第 1 行第 3 列，将指标序号填入表 9 第 1 行第 4 列。

指标体系 2: 由经典的 mRMR 方法[33]选出的指标体系。指标体系 2 共 23 个指标，将指标个数填入表 9 第 2 行第 3 列，将指标序号填入表 9 第 2 行第 4 列。

指标体系 3: 违约判别能力最强(即指标与违约状态之间互信息值最大)的 5 个指标构成的指标体系。根据表 7 第 45 行指标与违约状态之间的互信息值，从中选出互信息值最大的 5 个指标，将指标个数将指标个数填入表 9 第 3 行第 3 列，将指标序号填入表 9 第 3 行第 4 列。

指标体系 4: 由基于 SVM 向后次序选择指标遴选方法[39]筛选出的 11 个指标构成的指标体系。将指标个数将指标个数填入表 9 第 4 行第 3 列，将指标序号填入表 9 第 4 行第 4 列。

指标体系 5: 由基于 KNN 向后次序选择指标遴选方法[39]筛选出的 13 个指标构成的指标体系。将指标个数将指标个数填入表 9 第 5 行第 3 列，将指标序号填入表 9 第 5 行第 4 列。

指标体系 6: 所有 44 个指标构成的指标体系。将指标个数将指标个数 44 填入表 9 第 6 行第 3 列，将指标序号用“All”表示填入表 9 第 6 行第 4 列。

(2)指标体系 Φ 值的计算

本文用式(7)定义的最小冗余最大相关标准，即 Φ 值，来对比几个指标体系的优劣，Φ 值越大说明指标体系越能兼顾判别力和冗余度。

对于本文的指标体系 1，根据表 3 第 *e* 列的 0-1 向量，得到指标组合对应的决策向量 $c^{(1)}$，将 $c^{(1)}$代入上文式(16)中计算对应的 Φ 值，可得 $\Phi(c^{(1)})$=0.642，填入表 9 第 1 行第 5 列中。

根据表 9 第 4 列中各指标体系的指标序号，容易得到指标体系 2 至指标体系 6 的决策向量 $c^{(2)}$、$c^{(3)}$、$c^{(4)}$、$c^{(5)}$、$c^{(6)}$。分别将指标体系 2 至指标体系 7 的决策向量 $c^{(2)}$、$c^{(3)}$、$c^{(4)}$、$c^{(5)}$、$c^{(6)}$代入式(16)中，可以计算出指标体系 2 至指标体系 6 的 Φ 值：$\Phi(c^{(2)})$=0.141，$\Phi(c^{(3)})$=0.074，$\Phi(c^{(4)})$=0.024，$\Phi(c^{(5)})$=0.052，$\Phi(c^{(6)})$=0.071。并分别填入表 9 第 5 列第 2-6 行。

表 9 指标组合对比分析

Tab.9 The comparison between different indicator groups

(1)序号	(2)指标遴选模型	(3)指标数	(4)指标序号	(5)Φ值
1	本文模型	5	8、10、12、18、32	0.642
2	经典 mRMR 模型[33]	23	1、3、4、5、8、9、10、11、12、14、17、18、19、20、25、26、27、29、30、31、32、36、43	0.141
3	前 5 互信息值对应指标	5	11、12、14、19、43	0.074
4	基于 SVM 向后次序选择[39]	11	5、8、9、13、21、22、23、24、35、37、44	0.024
5	基于 KNN 向后次序选择[39]	13	13、14、15、17、18、19、22、23、26、27、29、40、43	0.052
6	所有指标	44	All	0.071

(3)指标体系对比

由表 9 第 5 列 Φ 值对比可知，本文所建立指标

体系对应最小冗余最大相关标准的Φ值最大，为0.642，优于经典的mRMR模型遴选的23个指标构成的指标组合的，也优于其他指标筛选方法得到的指标体系。

通过6个指标体系的对比可以看出，本模型筛选出的指标体系相对其他指标体系既能保证具有较强的违约鉴别能力，又能保证足够精简，是基于mRMR 标准筛选出的中国农户最优信用评价指标体系。

4.7 两个命题的证明

(1)命题1及证明

命题1：由单个违约鉴别能力强的指标构成的指标组合，违约鉴别能力不一定强。

证明：根据上文4.6中的模型对比分析，表9第1行第5列本文遴选的5个指标构成的指标组合的$\Phi(c^{(1)})$=0.642远大于表9第3行第5列同样5个单个违约鉴别力大的指标组成的指标组合的$\Phi(c^{(3)})$=0.074，则可以证明单个违约鉴别力最强的指标构成的指标组合的整体违约鉴别能力不一定也强。

(2)命题2及证明

命题2：指标组合的个数不是越多越好。

证明：通过将上文式(17)进行修改，令约束条件分别为$\sum_{i=1}^{44} c_i = n$ (n=1,2,...,44)，依次控制指标组合的个数，每控制一次指标组合中指标的个数，就解一次规划方程，即可得到一个指标组合。因此可以得到指标个数从1到n的n个最优指标组合，计算这n个指标组合的Φ值，并绘制出折线图，如图3所示。图3的横坐标对应的是指标组合中指标个数，纵坐标是指标组合对应的Φ值，由图3可知，Φ值随着指标个数增加，呈现先增加后减小的趋势，因此可以说明指标组合的个数不是越多越好。

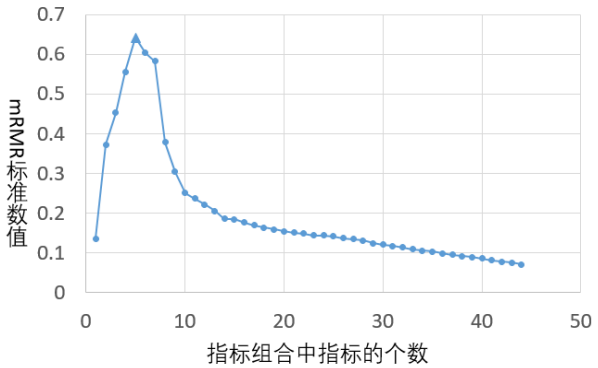

图3 44个指标组合的mRMR标准Φ值

Fig.3 Φ value of 44 indicator groups

5 结论

5.1 主要结论

(1)由于指标之间具有相关性，单个违约鉴别力强的指标构成的指标组合整体违约鉴别力不一定也强。由上文4.6的实证表明，“5个违约鉴别力最强的指标组成的体系的Φ值为0.074”远小于本文建立的体系的Φ值为0.642，说明单个指标鉴别能力最强，组合起来的体系鉴别能力反而不强。

(2)评级指标体系内指标个数并不是越多越好。根据上文“4.7”以及图3可知，依次固定指标个数得到的最优指标组合的Φ值随着指标个数的增加，先增加后减小，具有一个峰值。因此，对评级指标体系而言，并不是指标个数越多越好。

(3)本文指标遴选模型优于基于增量搜索方法的经典mRMR模型。根据上文表9中第1和2行第5列的Φ值对比，本文基于mRMR标准通过解整数规划遴选出来的5个指标构成的指标组合比经典的基于mRMR标准通过增量搜索算法[35]遴选出来的23个指标具有更高的Φ值，即本文基于整数规划的方法更优，能搜索出鉴别能力更高的指标体系。

(4)本文找出了影响中国农户信用状况的关键指标状态。

对中国农户贷款的实证研究表明，当家庭人数与劳动力人数的比值在3.5到4之间时，农户贷款更容易违约。可能是因为这种家庭人数与劳动力人数高比例的家庭结构，只有少数成员具有收入来源，在还款时存在困难。

居住年限在1到32年之间的农户信用水平相对于其他居住年限的农户更好，而居住少于1年的农户信用水平相对更差。可能是因为居住年限过短的客户稳定性较差，缺乏一定的还款能力。

当农户历史上申请贷款次数少于4次时，农户贷款违约的人数更多。从反面可以看出，申请贷款次数超过4次时的农户信用水平较高，说明好的信用记录能够延续。

5.2 主要创新

(1)以所有指标组合与违约状态之间的平均“互信息”最大为目标来保证遴选的指标组合具有最大的违约鉴别能力，以所有指标组合中的指标与指标之间的平均“互信息”最小为目标来保证在被遴选的指标组合中，指标间的信息冗余最小；通过这种指标组合最小冗余最大相关(mRMR)为目标建立了0-1规划模型、遴选出兼顾违约判别能力最大和冗余度最小的最优指标组合。通过0-1规划的全局最优解，改变了现有研究采用增量搜索算法仅仅会得到局部最优解的弊端。

(2)通过$\chi 2$统计量把最优指标组合中的每一个数值型指标划分成不同区间或状态，并用虚拟变量编码表示每一个数值型和类别型指标的状态，以所有指标的不同的状态(而不是对指标数值)为自变量，以客户的真实的违约状态为因变量进行逻辑回归，通过逻辑回归方程的系数β_i揭示指标组合中不同状态对违约鉴别能力的影响，完善了现有研究只遴选指标组合，忽略揭示指标状态对违约鉴别能力影响的问题。

参考文献

[1] Chandrashekar G, Sahin F. A survey on feature selection methods[J]. Computers & Electrical Engineering, 2014,

40(1):16-28.
[2] Piramuthu S. Feature Selection for Financial Credit-Risk Evaluation Decisions[J]. INFORMS, 1999,11(3):258-266
[3] Liu Y. Data Mining Feature Selection for Credit Scoring Models[J]. Journal of the Operational Research Society, 2005, 56(9):1099-1108.
[4] Aryuni M, Madyatmadja E D. Feature Selection in Credit Scoring Model for Credit Card Applicants in XYZ Bank: A Comparative Study[J]. International Journal of Multimedia & Ubiquitous Engineering, 2015, 10(5):17-24.
[5] Koutanaei F N, Sajedi H, Khanbabaei M. A hybrid data mining model of feature selection algorithms and ensemble learning classifiers for credit scoring[J]. Journal of Retailing & Consumer Services, 2015, 27:11-23.
[6] Chen Y S. Classifying credit ratings for Asian banks using integrating feature selection and the CPDA-based rough sets approach[J]. Knowledge-Based Systems, 2012, 26(1):259-270.
[7] Bouaguel W, Mufti G B, Limam M. Rank aggregation for filter feature selection in credit scoring[M]//Mining Intelligence and Knowledge Exploration. Springer, Cham, 2013: 7-15.
[8] Sadatrasoul S, Gholamian M, Shahanaghi K. Combination of Feature Selection and Optimized Fuzzy Apriori Rules: The Case of Credit Scoring[J]. International Arab Journal of Information Technology (IAJIT), 2015, 12(2).
[9] Dahiya S, Handa S, Singh N. A rank aggregation algorithm for ensemble of multiple feature selection techniques in credit risk evaluation[J]. International Journal of Advanced Research in Artificial Intelligence, 2016, 5: 1-8.
[10] Shian-Chang Huang, Ming-Hsiang Huang. Using SVMs with embedded recursive feature selections for credit rating forecasting[J]. Journal of Statistics & Management Systems, 2010, 13(1):165-177.
[11] Wang J, Hedar A R, Wang S, et al. Rough set and scatter search metaheuristic based feature selection for credit scoring[J]. Expert Systems with Applications, 2012, 39(6): 6123-6128.
[12] Wang J, Guo K, Wang S. Rough set and Tabu search based feature selection for credit scoring[J]. Procedia Computer Science, 2010, 1(1): 2425-2432.
[13] Wang C M, Huang Y F. Evolutionary-based feature selection approaches with new criteria for data mining: A case study of credit approval data[J]. Expert Systems with Applications, 2009, 36(3): 5900-5908.
[14] Hajek P, Michalak K. Feature selection in corporate credit rating prediction[J]. Knowledge-Based Systems, 2013, 51(1):72-84.
[15] Wang Q, Hu Y, Li J. Community-Based Feature Selection for Credit Card Default Prediction[C]// International Workshop on Complex Networks and their Applications. Springer, Cham, 2017:153-165.
[16] Maldonado S, Pérez J, Bravo C. Cost-based feature selection for Support Vector Machines: An application in credit scoring[J]. European Journal of Operational Research, 2017, 261(2): 656-665.
[17] Maldonado S, Bravo C, López J, Pérez, J. Integrated framework for profit-based feature selection and SVM classification in credit scoring[J]. Decision Support Systems, 2017, 104: 113-121.
[18] Chen F L, Li F C. Combination of feature selection approaches with SVM in credit scoring[J]. Expert Systems with Applications, 2010, 37(7):4902-4909.
[19] Bouaguel Waad, BelMufti Ghazi, Limam Mohamed. A three-stage feature selection using quadratic programming for credit scoring[J]. Applied Artificial Intelligence, 2013, 27(8):721-742.
[20] Oreski S, Oreski G. Genetic algorithm-based heuristic for feature selection in credit risk assessment[J]. Expert systems with applications, 2014, 41(4): 2052-2064.
[21] Stąpor K, Smolarczyk T, Fabian P. Heteroscedastic discriminant analysis combined with feature selection for credit scoring[J]. Statistics in Transition new series, 2016, 17(2): 265-280.
[22] Wang D, Zhang Z, Bai R, et al. A hybrid system with filter approach and multiple population genetic algorithm for feature selection in credit scoring[J]. Journal of Computational & Applied Mathematics, 2017.
[23] Altman EI. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy[J]. The Journal of Finance,1968, 23(4):589-609.
[24] Grablowsky BJ, Talley WK. Probit and discriminant functions for classifying credit applicants-a comparison[J]. Journal of Economics and Business, 1981 ,33(3):254-61.
[25] Boyes W J, Hoffman D L, Low S A. An econometric analysis of the bank credit scoring problem [J]. Journal of Econometrics, 1989, 40(1):3-14.
[26] Duca J V, Whitesell W C. Credit cards and money demand: A cross-sectional study[J]. Journal of Money, Credit and Banking, 1995, 27(2): 604-623.
[27] Hand D J, Henley W E. Statistical classification methods in consumer credit scoring: a review[J]. Journal of the Royal Statistical Society: Series A (Statistics in Society), 1997, 160(3):523-541.
[28] Dong G, Lai K K, Yen J. Credit scorecard based on logistic regression with random coefficients[J]. Procedia Computer Science, 2010, 1(1):2463-2468.
[29] West D. Neural network credit scoring models[J]. Computers & Operations Research,2000,27(11):1131-52.
[30] Tsai C F, Wu J W. Using neural network ensembles for bankruptcy prediction and credit scoring[J]. Expert Systems with Applications, 2008, 34(4):2639-2649.
[31] Ong CS, Huang JJ, Tzeng GH. Building credit scoring models using genetic programming[J]. Expert Systems with Applications. 2005,29(1):41-47.
[32] Abdou HA. Genetic programming for credit scoring: The case of Egyptian public sector banks[J]. Expert Systems with Applications,2009,36(9):11402-11417.
[33] Ding C, Peng H. Minimum Redundancy Feature Selection from Microarray Gene Expression Data[J]. Journal of Bioinformatics & Computational Biology, 2005, 3(02):185-205.
[34] Kerber R. ChiMerge: discretization of numeric attributes[C]. Tenth National Conference on Artificial Intelligence. AAAI Press, 1992:123-128.
[35] Peng H, Long F, Ding C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2005, 27(8):1226-1238.
[36] Thomas L C, Crook J, Edelman D. Credit Scoring and Its Applications[M].Credit scoring and its applications, Society for Industrial and Applied Mathematics, 2002.
[37] 迟国泰，石宝峰. 基于信用等级与违约损失率匹配的信用评级系统与方法[P].中国: ZL201210201461.6. 2015-08-19.
Chi Guotai, Shi Baofeng. Credit Rating System and Method Based on Matching of Credit Rating and LGD [P]. China: ZL201210201461.6. 2015-08-19.
[38] 迟国泰，程砚秋. 基于信用等级与违约损失率匹配的信用评级调整方法[P].中国: ZL 2012 1 0201114.3. 2015-11-18.

Chi Guotai, Cheng Yanqiu. Credit rating adjustment method based on Matching of Credit Rating and LGD[P].China: ZL 2012 1 0201114.3. 2015-11-18.
[39] Aha D W, Bankert R L. A Comparative Evaluation of Sequential Feature Selection Algorithms[J]. Lecture Notes in Statistics, 1996, 112:199-206.

Early Warning Research on Small Enterprise Credit Risk Based on Optimal Feature Subset of Big Data

Chi Guotai, Zhang Tong
(School of Economics and Management, Dalian University of Technology, Dalian, 116024)

Abstract: The credit rating is to evaluate the customer's credit level and help financial institutions to make loan decisions. This study focuses on two issues in credit rating, the first one is selecting a group of variables with strong default discriminal ability and low redundant, the second one is finding out the state of variables significantly influence customor's credit. **The innovations and features of this paper** are as follows. **Firstly**, we take the average mutual information for variable and default state as the correlation (also known as discriminal ability) of the variable group, and similarly take the average mutual information for variable and variable as the redundancy of variable group. With the with the minimum redundancy maximum correlation (mRMR) criteria, we construct a 0-1programming model to select an optimal variable group which improve the classical method based on incremental search. As a result, five best variables are selected which are labor member, familiy member/labor member, living period, the lender's household expenses and so on. **Secondly**, we convert the numerical variable into category variable through the ChiMerge binning method, and then encode the category variables to dummy variables considered as different states of the variables. After that, we use a logistic regression to fit the dummy variables and default state. By analyse the coefficient of the regression we conclude that when value of the variable 'the familiy member/labor member' are in the interval [3.5, 4), the farmer tends to default. More over, the farmer who borrow money less than 4 times from the bank are inclined to default. Also, the resident time influence the credit, the farmer living more than one year have higher credit level than those living less than one year.

Key Words: Credit rating; group variable selection; optimal variable group; states of variable; mutual information; 0-1 programming