基于相似样本归并的大样本混合信用评估模型①

张润驰，杜亚斌，薛立国，徐源浩，吴心弘
(南京大学商学院，南京 210093)

摘要： 当前面向大样本设计的信用评估模型，大多没有深入探究大样本的分布特征，只是简单地将传统评估方法应用在大样本上. 首先提出了用于描述大样本分布特征的相关属性集、边界向量等若干概念及定义，并证明了其主要性质. 之后在两个大样本数据集的基础上，研究了样本在相似性方面的分布特征，最后设计了一种大样本混合信用评估模型——HLSCE 模型. HLSCE 模型认为在大样本数据集中，样本的同一属性在不同局部区域内，对分类性能的贡献是不同的. 具体地，HLSCE 模型根据各样本与边界向量的相似性差异，结合生物启发式算法，将样本归并划分为若干子集并分别在其上训练基分类器. 实证研究表明，HLSCE 模型的分类精度相比于现有的代表性信用评估模型更高，同时也具有更为优越的平衡性与稳定性.

关键词： 信用风险；信用评估；大样本；边界向量
中图分类号： F83 **文献标识码：** A **文章编号：** 1007－9807(2018)07－0077－14

0 引 言

信贷过程中的信用风险是指获得信用支持的债务人不能遵照合约按时足额偿还本金和利息的可能性. 当代以银行业为代表的信贷机构一般选择以信用评估技术对信用风险进行预测与管理. 信用评估是一整套帮助贷款机构发放贷款的决策模型及其支持技术[1]. 随着现代商业银行的不断发展与壮大，各银行构建自己内部的信用评估模型渐成常态. 巴塞尔协议 II 亦明确指出，在满足某些最低条件和披露要求的前提下，有资格采用内部评级(IRB)法的银行可以根据自己对风险要素的估计值决定对特定暴露的资本要求.

就问题性质而言，信用评估在本质上是一种分类问题，根据待评估样本的多维度属性特征，设计合适的模型将所有样本分为若干类(如信用好和信用差两类). 现有的诸多模型根据使用技术的不同，大体可以分为基于传统统计理论的模型与基于数据挖掘技术的模型两类.

基于传统统计理论的模型，主要以统计学领域的统计推断方法为核心进行模型的构建. Orgler[2] 最初基于线性回归模型设计了商业贷款评分模型，但考虑到线性回归模型不能保证其因变量(违约概率)分布于 [0,1] 间，Steenackers 等人[3] 进一步提出了以逻辑回归(logistic)函数为核心的 logistic 评估模型，由于 logistic 模型具有结构简单、解释力强等特点，至今依然被广泛运用. Leonard 等[4] 将经典的贝叶斯原理引入信用风险管理模型，我国学者王春峰等人[5] 基于小样本 CV 方法对信用评估问题进行了深入研究，张洪祥等[6] 则在现有模型中引入了时间序列与灰色评价体系，进一步提高了模型的适用性.

此后，随着机器学习、人工智能等领域的逐渐发展，一批研究者尝试将数据挖掘方法应用于信

① 收稿日期：2016－03－21；修订日期：2017－10－29.
基金项目：国家自然科学基金重大研究计划资助项目(90718008)；江苏省自然科学基金资助项目(2004119).
作者简介：张润驰(1990—)，男，江苏南京人，博士生. Email：zhangrunchi@aliyun.com

用风险的管理研究中. Leong 等人[7-9]以人工神经网络(ANN)为工具设计了多种信用评估模型,Sermpinis 等研究者[10,11]引入或改进了支持向量机方法(SVM),取得了较为优越的评估性能. 庞素琳等[12]研究了 C5.0 决策树在个人信用评级中的应用,提高了模型的可解释性,Henley[13]则根据惰性学习理论设计了 k 最近邻(k-NN)模型. Kozeny[14]验证了遗传算法(GA)的不同目标函数选择对模型性能的影响,Hsieh 等[15-17]则基于现有的若干代表性模型的基本技术与特征,设计了多种混合信用评估模型.

但上述研究均是基于小样本展开,研究样本的规模一般不超过 1 000. Finlay[18,19]等研究者指出,基于小样本构建的模型容易对训练样本产生过度拟合,过度学习小样本的噪声特征,从而降低模型对训练集外未知样本的预测能力. 当前,以银行业为代表的信贷机构在经历了较长时间发展后,往往积累了大量的历史信贷记录;同时随着大数据时代[20,21]的到来,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,结合"互联网+"模式的大数据信用风险管理体系也逐渐成为研究热点,信用风险管理也逐渐进入"大样本"时代.

关于大样本的定义,学术界至今没有统一的标准. 在 Lehmann[22]经典的《Elements of Large-Sample Theory》一书中,大样本被定义为"当样本数 n 趋于无穷大时的样本规模". García[23]等近代研究者在设计大样本信用评估模型时,定义样本数小于 1 000 的为小样本,1 000 ~ 10 000 间为中样本,大于 10 000 则为大样本. 选择沿用 García 的定义,定义样本记录数大于 10 000 的样本为大样本.

近年来,一些学者开始基于大样本展开对信用评估模型的研究. Bellotti[24]等基于 25 000 份国外某金融机构样本,研究了 SVM 模型在大样本环境下的评估性能. Tsaih[25]等以台湾 41 000 家中小企业的数据为依据,设计了中小企业贷款评估模型,Harris[26]则基于 Barbados 某银行 1997 年—2012 年间的 21 620 组个人信贷样本,建立了基于聚类和 SVM 方法组合的信用评估模型. 我国学者方匡南等[27]也根据我国某商业银行 70 535 笔信贷记录,以随机森林为主要工具研究了信息不对称下的信用风险管理方法. Lessmanna[28]在公开的 PAKDD 与 kaggle 两个大样本数据集上研究了多种传统评估模型在大样本数据集上的性能.

然而,上述面向大样本的研究,大多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?这些正是值得研究的重点.

首先提出了关于大样本数据集分布特征的若干定义,证明了其主要性质;之后在两个大样本数据集的基础上,研究了大样本在相似性方面的分布特征;最后设计了一种根据样本间的相似性差异,结合生物启发式算法将完整的大样本数据集划分为近正类样本集、近负类样本集、中类样本集三个子集,在三个子集上分别构造分类器的大样本混合信用评估模型——HLSCE(hybrid large sample credit evaluation)模型. 实证研究结果表明,HLSCE 模型拥有更高的分类精度,也具有更为优越的平衡性与稳定性.

1 相关定义

首先给出下文用到的若干概念定义.

设大样本数据集共有 n 个样本 $X_i(i \in \{1,2,\cdots,n\})$,各样本包含 m 维属性 $A_j(j \in \{1,2,\cdots,m\})$. 各样本的类别标签 $Y_i \in \{0,1\}(i \in \{1,2,\cdots,n\})$. 若 $Y_i = 0$,则表示其对应的 X_i 为信用较好的负类(negative)样本;若 $Y_i = 1$,则表示其对应的 X_i 为信用较差的正类(positive)样本.

一般而言,不同样本在同一维属性上的取值会有所差异,即不存在 $\forall X_i = X_j(i,j \in \{1,2,\cdots,n\}, i \neq j)$ 的情况(若某一属性确实在所有样本中取值唯一,显然在模型中包含该属性并不能提高分类精度,现有的信用评估模型一般也会在预处理阶段剔除该属性). 因此可以给出模型唯一的

一个假设.

假设 1　样本数据集在每一维属性上的取值不唯一.

若存在函数 $Y=f(X)$，其度量了 n 个样本 X_i 与 Y_i 在均值意义上的相关关系，根据各属性 $A_j(j\in\{1,2,\cdots,m\})$ 在函数 $Y=f(X)$ 中系数 $p_j(j\in\{1,2,\cdots,m\})$ 的不同，有

定义 1　称包含最大数目属性 $A_j(j\in\{1,2,\cdots,m\})$ 的属性集 A^+ 为信用风险正相关属性集，若对 $\forall A_j\in A^+, p_j>0$. 同理，称包含最大数目属性 $A_j(j\in\{1,2,\cdots,m\})$ 的属性集 A^- 为信用风险负相关属性集，若对 $\forall A_j\in A^-, p_j<0$.

信用风险正相关属性集与信用风险负相关属性集度量了不同属性对信用风险的影响方向. 对于信用风险正相关属性集中的任一属性 $A_j(j\in\{1,2,\cdots,m\})$，其取值越大，在其他属性不变的条件下，该样本的信用风险越高. 同样地，对信用风险负相关属性集中的任一属性 $A_j(j\in\{1,2,\cdots,m\})$，其取值越大，在其他属性不变的条件下，该样本的信用风险越低.

定义 2　若存在 m 维样本 V，其在每一维上的分量 $V_k(k\in\{1,2,\cdots,m\})$ 满足

$$V_k=\begin{cases}\max(x_{ik}),\forall i\in\{1,2,\cdots,n\} & \text{if } A_k\in A^+\\ \min(x_{ik}),\forall i\in\{1,2,\cdots,n\} & \text{if } A_k\in A^-\end{cases} \tag{1}$$

则称该向量 V 为样本集的正类边界向量，简记为 V^+；类似地，若存在 m 维向量 V，在每一维上的分量 $V_k(k\in\{1,2,\cdots,m\})$ 满足

$$V_k=\begin{cases}\max(x_{ik}),\forall i\in\{1,2,\cdots,n\} & \text{if } A_k\in A^-\\ \min(x_{ik}),\forall i\in\{1,2,\cdots,n\} & \text{if } A_k\in A^+\end{cases} \tag{2}$$

则称该向量 V 为样本集的负类边界向量，简记为 V^-.

根据定义 2，V^+ 与 V^- 在各属性上的边界(最大或最小)取值可以取自同一类别中的不同样本. 事实上，V^+ 与 V^- 分别代表了根据样本集中各维属性与信用风险的关联特征构建的“信用风险最大的正类样本”与“信用风险最小的负类样本”. 易证 V^+ 与 V^- 具有如下几个性质.

性质 1　$V^+\neq V^-$

证明　运用反证法，若 $V^+=V^-$，有 $V_k^+=V_k^-(k\in\{1,2,\cdots,m\})$，根据定义 2，对 $\forall i\in\{1,2,\cdots,n\}$，$\forall k\in\{1,2,\cdots,m\}$，则有 $\min(x_{ik})=\max(x_{ik})$，即数据集在各个属性上的取值均唯一，这与假设 1 相违背，故 $V^+=V^-$ 的假设不成立，$V^+\neq V^-$.

Hsieh[15,26] 等研究者以 Euclidean 距离描述样本间的相似性，对任一对样本 X_a、X_b $(a,b\in\{1,2,\cdots,n\})$，其距离的计算公式为

$$dist(X_a,X_b)=\sum_{j=1}^{m}(X_{aj}-X_{bj})^2 \tag{3}$$

直观地，式(3)度量了两个样本 X_a、X_b 在 m 维属性上的平均差异程度，差异性越高，$dist(X_a,X_b)$ 的值越大，即两个样本的相似性越低. 但由于 Euclidean 距离其实隐含了样本服从正态分布的假设，而下一节的研究亦表明大样本数据集事实上并不服从正态分布，故选择以 Manhattan 距离为基础描述样本间的相似性. 此外，考虑到一般而言样本集的不同属性对信用风险的预测作用也各不相同，在度量样本相似性时，应当区分各属性贡献度的差异，故对各属性赋予权重 $w_j(j\in\{1,2,\cdots,m\})$，使得与信用风险关联较大的属性有较大的权重，从而在以距离度量相似性时，在信用风险关联较大的属性上，差异更大的样本之间距离更大. 加权 Manhattan 距离的计算公式为

$$dist(X_a,X_b)=\sum_{j=1}^{m}w_j\times|X_{aj}-X_{bj}| \tag{4}$$

结合式(4)，易证 V^+ 与 V^- 具有如下的性质 2.

性质 2　任一对样本 X_a、X_b $(a,b\in\{1,2,\cdots,n\})$ 间的相似性大于等于 V^+ 与 V^- 间的相似性，即对 $\forall a,b\in\{1,2,\cdots,n\}$，有 $dist(X_a,X_b)\leqslant dist(V^+,V^-)$.

证明　根据定义 2，有

$$\begin{aligned}dist(V^+,V^-)&=\sum_{j=1}^{m}w_j\times|\min(X_{ij})-\max(X_{ij})|\geqslant\\ &\sum_{j=1}^{m}w_j\times|X_{aj}-X_{bj}|\\ &=dist(X_a,X_b)\ (\forall i\in\{1,2,\cdots,n\})\end{aligned} \tag{5}$$

Bellotti[24] 等研究者指出，logistic 模型估计的各属性系数可以作为各属性与信用风险相关性的衡量标准. 因此在下文的研究中，统一在样本标准化的基础上，选择各属性在 logistic 模型中的参数正负作为信用风险正相关属性集与信用风险负相关属性集的划分标准，以各属性在 logistic 模型中的系数绝对值大小作为各属性的权值.

2 大样本数据集的分布特征

在上述定义的基础上，这一节首先通过分析两类样本与 V^+、V^- 在相似性方面的分布差异，研究大样本数据集的分布特征.

研究选用两组公共大样本信用评估数据集——PAKDD 数据集与 kaggle 数据集. PAKDD 数据集为 2010 年举办的亚太知识发现与数据挖掘会议提供的公开数据集，kaggle 数据集则取自数据科学领域著名的 kaggle 网络社区. Lessmann[1,28] 等人的研究均基于这两组数据集.

PAKDD 数据集共含有 50 000 条 53 维属性的记录，去除含缺失值的记录与含义不明、意义不大的属性(如 ID、邮编、电话区号等)与取值唯一的属性后，剩余 49 935 个 23 维样本，其中正类样本 13 018 个，负类样本 36 917 个，负类/正类样本比率为 2.84:1. 因变量属性为是否该接受该贷款申请，剩余 22 维的自变量属性分别为月常规收入、月平均其他收入、个人资产估价、填写地址是否为家庭地址、性别、是否有家庭电话、是否有邮箱、是否是 Visa 卡用户、是否是 Master 卡用户、是否是 Siners 卡用户、是否是 American Express 卡用户、是否是其他卡用户、是否有固定工作、是否有专业电话、申请者选择的每月还款日、家庭成员数目、房屋居住月数、银行普通账户数目、银行特殊账户数目、拥有汽车数目、当前工作持续的月数以及年龄.

kaggle 数据集共含有 150 000 条 10 维属性的记录，各属性含义明确，在去除含缺失值的记录后，剩余 120 269 个 11 维样本，其中正类样本 8 357个，负类样本 111 912 个，负类/正类样本比率为 13.39:1，存在较为严重的负类样本偏倚现象. 各样本的因变量属性为当前贷款是否逾期 90 天及以上，10 维自变量属性分别为信用额度的余额占信用额度总和、年龄、过去贷款逾期 30 天 ~ 59 天的次数、过去贷款逾期 60 天 ~ 89 天的次数、过去贷款逾期 90 天及以上的次数、每月债务与必须支出占总收入的比率、月收入、债务和信用卡总数、住房抵押贷款数以及家庭成员数目.

首先直接观察在两个数据集中，正类、负类样本在与 V^+、V^- 的相似性分布方面是否存在差异. 根据性质 2，V^+ 与 V^- 间的距离为任一对样本间的最大可能距离. 故选择将 V^+ 与 V^- 间的距离划分为 100 个均等距离区间，分别统计正类样本与 V^+、V^- 的距离、负类样本与 V^+、V^- 的距离在各个区间中的分布情况，其结果如图 1 所示. 图中 X 轴左侧原点代表与 V^+、V^- 距离为 0(即最为相似)的对应样本，X 轴代表了划分的各距离区间，Y 轴为该区间中的两类样本数占各自同类总样本数的百分比.

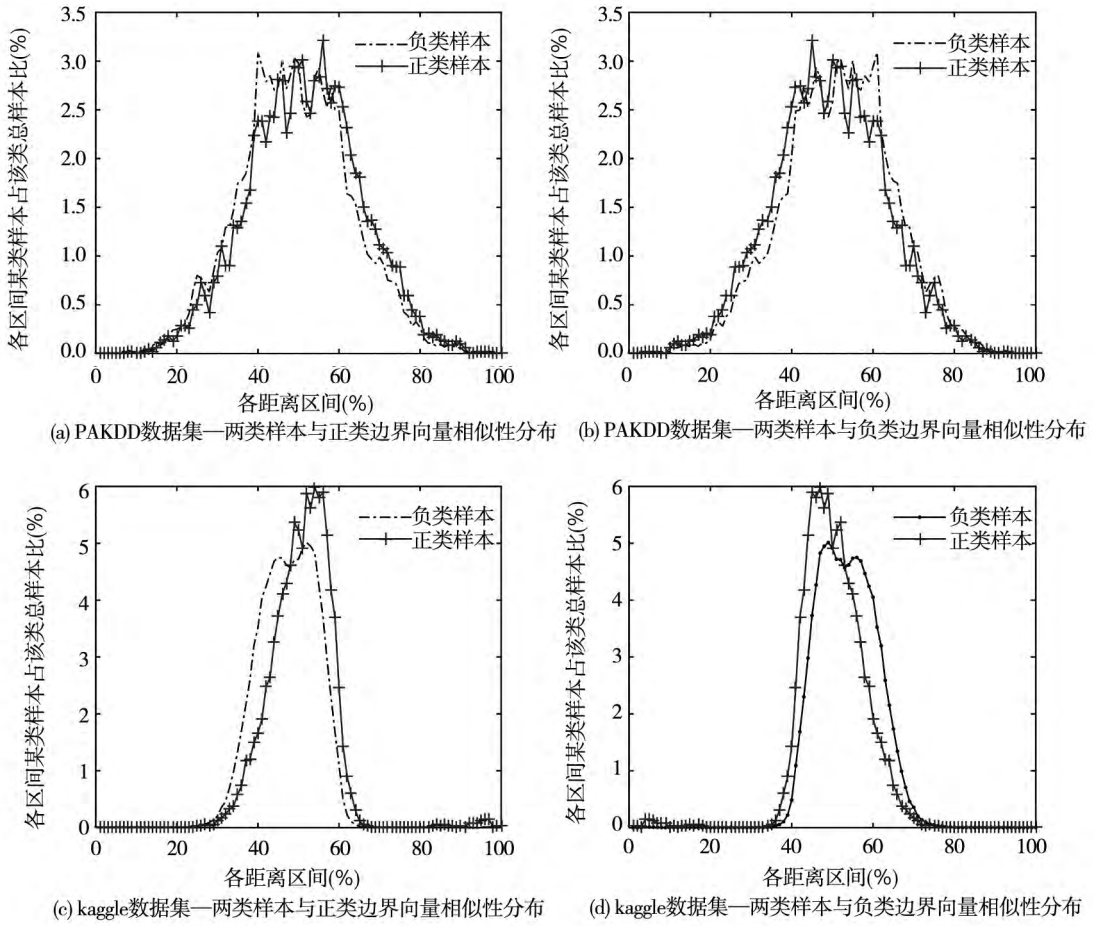
从图 1 中可以直观地看出，两类样本与 V^+、V^- 间的距离分布具有类似钟形曲线的特征. 在 PAKDD 数据集中，负类样本的分布相比于正类样本更加相似于 V^+(表现为在图 1. a 的前 50% 距离区间内，负类样本的分布占比高于正类样本)，同时正类样本的分布更加相似于 V^-(表现为在图 1. b 的后 50% 距离区间内，正类样本的分布占比高于负类样本). 在 kaggle 数据集中，两类样本的总体分布情况与 PAKDD 数据集相比显得更加趋于中心化，样本主要集中在 40% ~ 60% 的距离区间内，同时负类样本的分布相比于正类样本更加相似于 V^+，而正类样本则相反.

从实际信用风险管理角度出发，信贷机构在面对一个多数指标都较差的贷款申请者时，往往会判断其信用风险较高，从而出于谨慎拒绝贷款，因此多数指标较差、可能具有较高信用风险的申请者，最终并未成为信贷机构的客户，因此其样本也没有被纳入发放贷款客户的样本集. 而存在于样本集中的违约个体，在贷款申请之初往往具有与未违约贷款者相似的特征，从而被信贷机构认可其信用风险水平并发放贷款，这在一定程度上

解释了正类样本的分布与 V^- 更为接近的现象.

接下来，进一步研究正类、负类样本与 V^+、V^- 的相似性分布特征. 两类样本与 V^+、V^- 的相似性是否有可能服从同一分布？首先对两个数据集中两类样本与 V^+、V^- 的距离序列进行无需预先假设分布特征的两样本 Kolmogorov-Smirnov 检验，其检验结果如表 1 所示，两个数据集均在 1% 的显著性水平上拒绝了服从同一未知分布的原假设.

在社会科学领域，许多大样本数据集具有正态分布的特征，进一步检验两个数据集中各正类样本与 V^+，V^- 的距离、负类样本与 V^+，V^- 的距离是否服从正态分布. 首先假设上述四类距离均服从均值、方差未知的正态分布，对这四种假设进行 Jarque-Bera 检验与 Lilliefors 检验，其检验结果如表 2 所示，两个数据集上除正类样本与 V^+ 的距离分布、正类样本与 V^- 的距离分布在 5% 显著性水平上拒绝了原假设外，其余检验均在 1% 的显著性水平上拒绝了原假设.



(a) PAKDD数据集—两类样本与正类边界向量相似性分布　(b) PAKDD数据集—两类样本与负类边界向量相似性分布

(c) kaggle数据集—两类样本与正类边界向量相似性分布　(d) kaggle数据集—两类样本与负类边界向量相似性分布

图 1　正类、负类样本与 V^+、V^- 的距离分布

Fig. 1 Distance distribution between positive samples, negative samples and V^+, V^-

表 1　两个数据集上的同分布检验

Table 1 Test of same distribution on two data sets

	*KS*2 统计量	*p* 值	原假设	是否拒绝
PAKDD 数据集	0.080 2	2.663 2e－54	两类样本与 V^+ 的距离服从同分布	是
	0.080 2	2.663 2e－54	两类样本与 V^- 的距离服从同分布	是
kaggle 数据集	0.186 1	6.129 6e－235	两类样本与 V^+ 的距离服从同分布	是
	0.186 1	6.129 6e－235	两类样本与 V^- 的距离服从同分布	是

表 2 两个数据集上的正态分布检验

Table 2 Test of normality distribution on two data sets

	Jarque – Bera 检验				Lilliefors 检验			
	统计量	p 值	原假设	是否拒绝	统计量	p 值	原假设	是否拒绝
PAKDD 数据集	8.095 1	0.017 9	正类样本与 V^+ 距离服从正态分布	是	0.008 9	0.015 6	正类样本与 V^+ 距离服从正态分布	是
	8.095 1	0.017 9	正类样本与 V^- 距离服从正态分布	是	0.008 926 784	0.015 6	正类样本与 V^- 距离服从正态分布	是
	57.261 8	1.00e –03	负类样本与 V^+ 距离服从正态分布	是	0.013	1.00e –03	负类样本与 V^+ 距离服从正态分布	是
	57.261 8	1.00e –03	负类样本与 V^- 距离服从正态分布	是	0.013	1.00e –03	负类样本与 V^- 距离服从正态分布	是
kaggle 数据集	1.994e +04	1.00e –03	正类样本与 V^+ 距离服从正态分布	是	1.993 9e +04	1.00e –03	正类样本与 V^+ 距离服从正态分布	是
	1.994e +04	1.00e –03	正类样本与 V^- 距离服从正态分布	是	1.993 9e +04	1.00e –03	正类样本与 V^- 距离服从正态分布	是
	1.661e +03	1.00e –03	负类样本与 V^+ 距离服从正态分布	是	1.660 9e +03	1.00e –03	负类样本与 V^+ 距离服从正态分布	是
	1.661e +06	1.00e –03	负类样本与 V^- 距离服从正态分布	是	1.660 9e +03	1.00e –03	负类样本与 V^- 距离服从正态分布	是

综上，在两个数据集上，两类样本与 V^+、V^- 的距离（即相似性）分布均具有明显的钟形曲线分布特征，但两类样本与 V^+、V^- 的相似性并不服从同一分布，且均不服从正态分布.

3 基于大样本的信用评估模型

基于上一节的实证结论，这一节设计基于大样本的信用评估模型——HLSCE 模型. HLSCE 模型的核心思想是，在大样本数据集中，样本的同一属性在不同局部区域内，对分类性能的贡献是不同的. 因此，模型尝试根据各样本与前文定义的 V^+、V^- 之间的相似性，将各样本划分入恰当的子集，并在各子集上分别构建基分类器.

根据第一节的相关定义，HLSCE 模型以样本集在 logistic 模型中的参数正负为依据确定 V^+ 与 V^-，以参数大小为依据选择权重. 同时，考虑到在大样本数据集中可能含有一些冗余属性影响模型性能，故在样本预处理阶段，统一剔除在 logistic 模型中系数不显著（显著性水平低于 5%）的属性.

根据现有模型的通用做法，HLSCE 模型首先对样本集进行归一化. 归一化操作能统一不同属性的量纲，使属性间的绝对关系变为相对关系，从而使各属性能在同一取值区间中比较. 对式(4)而言，量纲的一致性使得在计算距离时不会出现因某一属性的量纲过大导致该属性成为影响距离的唯一主要属性的情况. 归一化过程为

$$X'_{ij} = \frac{X_{ij} - \min\left(X_{*j}\right)}{\max\left(X_{*j}\right) - \min\left(X_{*j}\right)} \tag{6}$$

其中 X_{ij}' 为第 i 个样本在归一化后的第 j 个属性值，X_{ij} 表示第 i 个样本在归一化前的第 j 个属性，X_{*j} 为所有样本在归一化前的第 j 个属性上的取值向量.

为便于进一步说明 HLSCE 模型的理论基础，首先在二维空间中绘出大样本相似性分布的示意图，如图 2 所示. 图 2 中圆形标记点代表负类样本向量，叉状标记点代表正类样本向量. 二维样本空间左上角为负类边界向量 V^-，右下角为正类边界向量 V^+，$dist(V^+, V^-)$ 即对角线距离，符合性质 2，各类样本与 V^+、V^- 的距离符合钟形曲线分布特征.

在构建基于大样本的分类模型时，一个重要问题是如何识别相似异类样本间的特征差异. Bellotti 等研究者[24-28]在构建模型时，直接将所有样本置入同一分类器进行训练，但由于大样本

数据集在不同局部区域中的数据特征往往存在差异,基于全样本训练的单一分类器只能在均值意义上对样本的可分类特征进行浅层识别,难以更好地利用数据集的局部特征. 考虑到各类样本与 V^+、V^- 的相似性特征,定义两个参数 k_{pos} 与 k_{neg} (k_{pos},k_{neg} ∈ (0,1)) ,分别表示近 V^+ 样本集与近 V^- 样本集中所含样本数占总样本数的比例,进而根据各样本与 V^+、V^- 的距离差异,将整体样本集划分为 3 个子集.

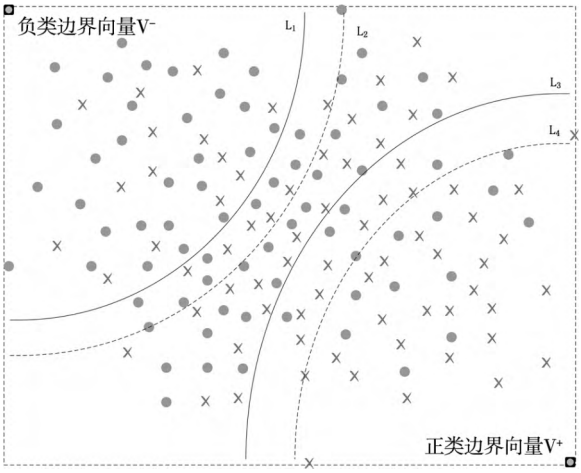


图 2　二维空间中大样本相似性分布示意图

Fig. 2 Sketch map of large sample similarity distribution in two dimensional space

近 V^+ 样本集,由与 V^+ 距离最近的 $k_{pos} \times n$ 个两类样本的并集构成(在图 2 中可由 V^+ 与 L_3 曲线、L_4 曲线构成的区域并集表示) .

近 V^- 样本集,由与 V^- 距离最近的 $k_{neg} \times n$ 个两类样本的并集构成(在图 2 中可由 V^- 与 L_1 曲线、L_2 曲线构成的区域并集表示) .

中类样本集,由所有不属于近 V^+ 样本集与近 V^- 样本集的两类样本构成(在图 2 中可由 L_2 曲线与 L_3 曲线间的区域表示) .

具体的划分方法为,首先根据式(4) 计算出两类样本与 V^+、V^- 的距离,之后对两类样本与 V^+、V^- 的距离序列分别按距离大小进行升序排序(即与 V^+、V^- 距离越近的样本,排序越前) ,在排序后的两类样本与 V^+ 的距离序列中,取前 $k_{pos} \times n$ 个样本构成近 V^+ 样本集,同理在排序后的两类样本与 V^- 的距离序列中,取前 $k_{neg} \times n$ 个样本构成近 V^- 样本集. 最后,对于剩余的所有未被近 V^- 样本集或近 V^+ 样本集包含的样本,将其置入中类样本集.

三个子集在样本的相似性上存在着明显的区别,近 V^+ 样本集含有最接近于 V^+ 的两类样本共 $k_{pos} \times n$ 个,近 V^- 样本集含有最接近于 V^- 的两类样本共 $k_{neg} \times n$ 个,而中类样本集则包含了与 V^+、V^- 相似性均不高的剩余所有样本.

根据两类样本与 V^+、V^- 的相似程度不同,将最为相似的若干两类样本构成一个集合,在该集合上训练模型,能够更好地利用异类相似样本间的可分类特征,从而使模型分类精度更高; 同时对于那些与 V^+、V^- 相似性皆不高的“模糊样本”,以单独构建中类样本集的形式避免将其硬性划入近 V^+ 样本集或近 V^- 样本集,维持各子集中样本间的相似性.

显然,k_{pos} 与 k_{neg} 的取值直接关系到三个样本子集的划分方式,并进而可能影响 HLSCE 模型的整体分类精度. 然而由于缺乏关于最优子集划分方式的先验信息,难以直接给出最优的 k_{pos} 与 k_{neg} 取值. 但考虑到划分参数的选择在本质上构成了一个组合优化问题,选择较优的 k_{pos} 与 k_{neg} 取值组合,使得 HLSCE 模型的最终预测精度最高,因此可以借助较适合解决组合优化问题的生物启发式算法对 k_{pos} 与 k_{neg} 的参数设定进行寻优. Kar 等人[29] 总结了目前研究最为成熟的 12 个生物启发式算法,综合算法成熟度、时间复杂度、对参数的敏感性等多方面因素,选择布谷鸟搜索(cuckoo search) 算法,对 HLSCE 模型中 k_{pos} 与 k_{neg} 的取值进行优化.

作为一种新颖的生物启发式算法,CS 算法最初由 Yang 等人[30] 提出,算法模仿了布谷鸟不断寻找更好的鸟巢进行巢寄生繁殖的机理,在每一次迭代中,若干布谷鸟(个体) 在固定数量的鸟巢(解空间) 中以 Lévy 飞行的方式随机选择新的鸟巢产蛋(可行解) ,并与之前的鸟巢进行对比,若新鸟巢的位置(适应度) 优于之前鸟巢,则用新鸟巢替代之前鸟巢,同时最好的鸟巢(当前最优解) 将延续至下一代,此外布谷鸟的蛋有 $P_a \in [0,1]$ 的概率被宿主鸟发现,一旦被发现,宿主鸟将重新选择鸟巢或摧毁布谷鸟的蛋(避免陷入局部最优

空间).

在 k_{pos} 与 k_{neg} 的参数优化阶段,CS 算法设置每个个体的维度(问题的规模,即 k_{pos} 与 k_{neg} 的可选取值)为 2,个体数目为 g,宿主鸟察觉概率 P_a,解的限制条件为 $0 < k_{pos} < 1$,$0 < k_{neg} < 1$,且 $k_{pos} + k_{neg} < 1$. 设训练样本集的总样本数为 n_{train},近 V^+ 样本集、近 V^- 样本集与中类样本集上的子集分类正确率分别为 C_{pos}、C_{neg} 与 C_{mid},则 CS 算法的目标函数为寻找最优的 k_{pos}、k_{neg} 组合,最大化模型基于三个子集的总体分类正确率,即最大化

$$k_{pos} \times C_{pos} + k_{neg} \times C_{neg} + (1 - k_{pos} - k_{neg}) \times C_{mid} \tag{7}$$

设最优组合为 k'_{pos}、k'_{neg},则其应当满足对于 $\forall k_{pos}, k_{neg}$,有

$$k_{pos} \times C_{pos} + k_{neg} \times C_{neg} + (1 - k_{pos} - k_{neg}) \times C_{mid} \leqslant k'_{pos} \times C_{pos} + k'_{neg} \times C_{neg} + (1 - k'_{pos} - k'_{neg}) \times C_{mid} \tag{8}$$

HLSCE 模型进而于训练阶段,根据 k_{pos}、k_{neg} 将总体训练样本集划分为三个子集,并分别训练基分类器,同时记录在近 V^+ 样本集中与 V^+ 相似性最弱的样本与 V^+ 的距离,以及近 V^- 样本集中与 V^- 相似性最弱的样本与 V^- 的距离. 在测试阶段,需要将每一个新样本恰当地分配给训练阶段基于三个子集构筑的三个分类器之一进行预测,具体做法为首先测算该样本与 V^+、V^- 的距离,即判断其与 V^+、V^- 的相似性,若该样本与 V^- 的距离小于其与 V^+ 的距离,则该样本不会被划入近 V^+ 样本集,但其可能应当被划入近 V^- 样本集,也可能应当被划入中类样本集,因此进一步比较该样本与 V^- 的距离、训练阶段记录的近 V^- 样本集中与 V^- 相似性最弱的样本与 V^- 的距离之间的大小,若前者小于后者,则将其划入近 V^- 样本集,否则划入中类样本集. 对于样本与 V^+ 的距离小于其与 V^- 的距离的情况亦是同理. 在各测试样本的应属子集划分完毕后,以训练阶段在该子集上训练的基分类器识别该样本所属类别. 模型在训练阶段与测试阶段的主要流程如图 3 所示.

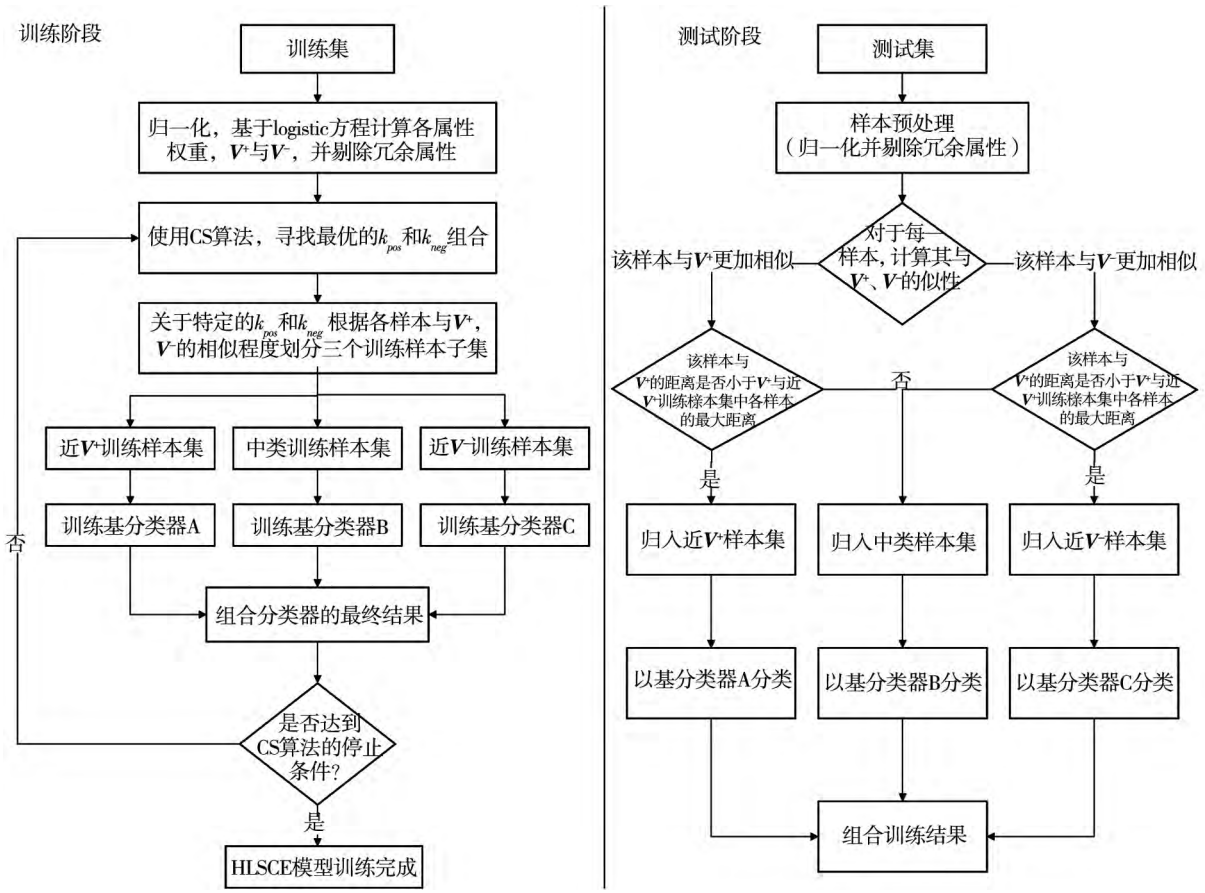


图 3　HLSCE 模型在训练阶段与测试阶段的主要流程

Fig. 3 The main procedure of HLSCE model in training stage and testing stage

面向大样本的信用评估模型除预测精度外，另一个值得关注的因素是时间复杂度，过高的时间复杂度会使模型的实用价值较低. 结合图 3，进一步分析模型的时间复杂度. 在训练阶段，归一化的过程需要对 n 个样本的 m 维属性进行操作，其时间复杂度为 $O(m \times n)$. 之后根据 logistic 方程计算各属性的权重并确定 V^+ 、V^- . logistic 方程以最大似然估计法求取参数，而最大似然估计法的估计过程一般以梯度下降法迭代求取，其时间复杂度为线性于问题维度的 $O(m \times n)$. 由于样本首先进行归一化操作，因此之后可直接根据各属性在 logistic 方程结果中的系数确定 V^+ 、V^- ，该步骤的时间复杂度为 $O(1)$. 接着基于 CS 算法优化 k_{neg} 与 k_{pos} 的参数组合，CS 算法本身是线性的，其具体时间复杂度取决于优化对象的时间复杂度，在 CS 算法内部需要对每一个样本与 V^+ 、V^- 进行相似度比较并划分三个训练子集，这一步骤的时间复杂度为 $O(m \times n)$ ，进而对两类样本与 V^+ 、V^- 的距离进行排序，采用快速排序算法的时间复杂度为 $O(n \times \ln\ n)$. 最后结合各基分类器进行模型训练，假设基分类器的时间复杂度为 $O(x)$ ，在设定的 CS 算法最大迭代次数 T 和个体数 g 的乘积远小于大样本的规模 n 、且样本的属性维度 m 远小于样本规模 n 时，CS 算法优化 k_{pos} 与 k_{neg} 参数组合部分的总体时间复杂度为

$$\begin{aligned}&\max\ (O(T \times m \times n), O(T \times n \times \ln\ n), \\ &\qquad O(T \times g \times x)) \qquad (9)\\ &\approx \max\ (O(n \times \ln\ n), O(x))\end{aligned}$$

由于以上各步骤顺序执行，因此 HLSCE 模型在训练阶段的总体时间复杂度为

$$\begin{aligned}&\max\ (O(m \times n), O(m \times n), O(1), \\ &\max\ (O(n \times \ln\ n), O(x))) \qquad (10)\end{aligned}$$

鉴于不同基分类器的时间复杂度各不相同，因此 HLSCE 模型的时间复杂度取决于基分类器与其它各步骤相比的时间复杂度相对量级，当基分类器的时间复杂度 $O(x)$ 大于 $O(n \times \ln\ n)$ 时，模型的时间复杂度为所选用的基分类器的时间复杂度; 当基分类器的时间复杂度小于等于 $O(n \times \ln\ n)$ 时，模型的整体时间复杂度为 $O(n \times \ln\ n)$. 在测试(实际应用) 阶段的时间复杂度不会超过训练阶段.

4　实证研究

为了充分检验 HLSCE 模型在结合不同基分类器情况下的性能，分别选择当前信用评估领域若干成熟的基分类器进行实证对比. 参与对比的基分类器为决策树(DT) 、判别分析(DA) 、逻辑回归(logistic) 、线性核的支持向量机(SVM-liner) 、高斯核的支持向量机(SVM-gauss) 、k 最近邻(k-NN) 与朴素贝叶斯(NB) ，这 7 种模型在现有信用评估研究文献中被大量使用，有较好的代表性.

在参数设置方面，k_{pos}、k_{neg} 的取值经 CS 算法优化确定，CS 算法的个体数设置为 15，初始化时每个个体随机生成，最大迭代次数设置为 30，算法停止条件为达到最大迭代次数.

在样本抽样方面，当前信用风险评估模型的研究文献大多采用 N 折交叉法对模型的性能进行检验. N 折交叉法将数据集均等地分为 N 个子集，每次迭代选取其中一个子集为测试集，其余 $N-1$ 个子集为训练集，其特点在于能使每一个样本在实验过程中均可参与训练过程或测试过程，但 Lee 等人[31] 的研究表明，由于在现实样本集中负类样本数总是多于正类样本数，而 N 折交叉法易受两类样本在总样本中的分布影响，可能因某一子集含有的两类样本分布不均造成模型的平衡性较差，Harris 等人[26,27] 因而采用按两类样本 1:1 比例抽样的 bootstrapping 抽样方法选择样本，但这种方法无法保证测试集和训练集之间的独立性，实证结果的稳健性不高. 为解决这一问题，首先使用 Chawla[32] 设计的 SMOTE 算法，生成新的少数类(正类) 样本，使得两类样本的比例为 1:1，之后采用 5 折交叉法进行实证，从而在确保训练集和测试集的独立性之余，避免因两类样本数量的悬殊差异导致各模型的平衡性较弱. 同时，为了保证对比的公允性和有效性，统一于样本预处理阶段剔除在 logistic 模型中系数不显著(显著性水平低于 5%) 的属性，之后基于相同的、经过精简的数据集检验 HLSCE 模型和对比模型的相对性能.

在模型性能评估方面，选用总体分类正确率、

正类分类正确率、负类分类正确率、正类召回率、负类召回率五个评价标准. 结合表 3 的混淆矩阵，这五个评价标准可分别定义为，总体分类正确率 $= \frac{TP + TN}{TP + FN + FP + TN}$，正类分类正确率 $= \frac{TP}{TP + FN}$，负类分类正确率 $= \frac{TN}{FP + TN}$，正类召回率 $= \frac{TP}{TP + FP}$，负类召回率 $= \frac{FN}{FN + TN}$.

表 3 混淆矩阵

Table 3 Confusion matrix

	实际正类	实际负类	总计
预测正类	TP	FN	$TP + FN$
预测负类	FP	TN	$FP + TN$
总计	$TP + FP$	$FN + TN$	$TP + FN + FP + TN$

除此之外，还需考察各个模型的平衡性与稳定性. 平衡性是指模型的分类能力在各类样本间较为平均，主要体现为正/负类分类正确率、正/负类召回率间的差异不大. 就实际的信用评估问题而言，模型一般容易对负类样本产生偏倚，即更有可能将正类样本识别为负类，而其代价(贷款损失)远比将负类样本误分为正类(失去潜在客户)要高，因此好的模型应当避免对负类样本的偏倚. 稳定性是指模型在测试阶段的性能与训练阶段相比变化不明显. 在实际应用中选择模型时，仅能根据模型基于历史数据的表现对模型未来的运行情况进行估计，因此风险厌恶的信用风险管理者往往更偏爱性能稳定的模型.

表 4 与表 5 分别展示了 HLSCE 模型结合 7 种基分类器，在 PAKDD 数据集与 kaggle 数据集上五个性能指标方面的对比情况. 其中，每一指标的取值对应两行，第一行为 HLSCE 模型将该分类器作为基分类器后的指标，第二行为仅使用该基分类器时的指标. 表 6 展示了两个数据集上，HLSCE 模型在参数寻优阶段确定的 k_{pos}、k_{neg} 组合，以及在寻优过程中所尝试的最优参数组合与最差参数组合在模型总体分类正确率方面的差异情况.

表 4 HLSCE 模型在 PAKDD 数据集上的性能

Table 4 Performance comparison of HLSCE model on PAKDD data set

检验阶段	指标	HLSCE (DT)	HLSCE (DA)	HLSCE (logistic)	HLSCE (SVM - liner)	HLSCE (SVM - gauss)	HLSCE (k - NN)	HLSCE (NB)
训练阶段	总体分类正确率(%)	76.50	62.58	60.25	55.75	58.38	61.85	81.36
		76.00	60.83	58.63	55.15	58.03	61.33	79.30
	正类分类正确率(%)	82.00	57.16	59.87	53.70	59.20	62.63	89.25
		78.76	60.16	58.03	54.90	56.42	62.40	91.11
	负类分类正确率(%)	71.66	68.36	60.26	56.51	57.54	60.78	76.40
		73.86	61.60	59.33	56.39	60.94	60.57	72.86
	正类召回率(%)	72.53	69.99	61.55	53.60	69.47	53.64	72.46
		70.75	64.30	62.35	61.20	71.15	57.45	65.10
	负类召回率(%)	77.43	46.41	55.33	48.18	31.24	67.12	88.29
		81.25	57.35	54.90	49.10	44.90	65.20	93.50
测试阶段	总体分类正确率(%)	57.18	62.63	58.92	54.78	55.54	60.92	74.81
		55.60	59.40	55.90	53.60	56.50	60.10	73.30
	正类分类正确率(%)	60.10	67.63	51.00	52.46	49.43	61.23	75.64
		57.02	58.84	55.45	52.78	55.15	61.19	73.15
	负类分类正确率(%)	51.78	56.25	60.61	56.76	57.79	60.20	72.81
		54.69	60.12	56.49	55.40	59.00	59.61	76.46
	正类召回率(%)	47.21	44.89	62.40	53.41	69.19	49.46	67.46
		46.60	63.20	60.00	60.20	69.60	56.60	66.60
	负类召回率(%)	61.31	57.14	63.70	48.93	42.09	61.14	76.45
		64.60	55.60	51.80	47.00	43.40	63.60	88.00

表 5　HLSCE 模型在 kaggle 数据集上的性能比较

Table 5 Performance comparison of HLSCE model on kaggle data set

检验阶段	指标	HLSCE(DT)	HLSCE(DA)	HLSCE(logistic)	HLSCE(SVM-liner)	HLSCE(SVM-gauss)	HLSCE(k-NN)	HLSCE(NB)
训练阶段	总体分类正确率(%)	91.85	70.33	76.53	59.26	63.62	67.85	80.65
		91.70	63.28	75.60	59.15	56.08	67.35	79.85
	正类分类正确率(%)	92.65	72.19	83.56	60.75	64.34	75.05	97.21
		92.12	63.70	81.72	57.90	68.63	71.68	99.13
	负类分类正确率(%)	90.26	68.71	70.84	58.51	59.83	59.62	72.14
		91.31	65.88	71.61	60.88	57.03	64.92	71.57
	正类召回率(%)	92.76	57.64	67.07	62.25	53.98	54.88	63.05
		91.20	61.65	66.15	67.15	60.35	59.10	60.20
	负类召回率(%)	91.60	70.23	84.88	57.75	60.02	70.19	97.16
		92.20	64.90	85.05	51.15	71.80	75.60	99.50
测试阶段	总体分类正确率(%)	78.94	65.68	72.82	58.36	56.91	66.80	71.96
		75.50	60.70	67.40	56.70	54.50	65.20	70.20
	正类分类正确率(%)	73.01	68.69	76.04	60.27	62.29	75.33	64.41
		75.30	61.18	72.59	55.45	67.01	69.52	62.00
	负类分类正确率(%)	80.91	60.30	60.57	56.09	52.03	60.15	73.88
		75.97	60.33	63.87	58.49	55.18	62.72	70.10
	正类召回率(%)	84.19	57.15	73.24	60.25	51.98	46.25	67.48
		76.20	59.50	58.80	64.60	38.20	57.60	61.60
	负类召回率(%)	78.88	77.47	71.65	55.75	64.02	72.57	74.11
		74.80	61.90	76.00	48.80	70.80	72.80	78.80

表 6　不同参数组合下 HLSCE 模型的性能比较

Table 6 Performance comparison of HLSCE model under different combinations of parameters

	PAKDD 数据集			kaggle 数据集		
分类器	kneg	kpos	最优与最差 kneg 与 kpos 组合下的性能差异(%)	kneg	kpos	最优与最差 kneg 与 kpos 组合下的性能差异(%)
HLSCE(DT)	0.50	0.26	4.13	0.10	0.27	1.94
HLSCE(DA)	0.23	0.38	3.20	0.24	0.10	8.33
HLSCE(logistic)	0.46	0.21	2.98	0.17	0.52	5.08
HLSCE(SVM-liner)	0.31	0.51	3.05	0.26	0.37	7.71
HLSCE(SVM-gauss)	0.28	0.38	1.27	0.42	0.35	7.55
HLSCE(k-NN)	0.42	0.15	8.53	0.65	0.12	8.36
HLSCE(NB)	0.28	0.41	4.09	0.45	0.12	8.40

从表 4 可以看出,在 PAKDD 数据集上的训练阶段,HLSCE 模型在结合各分类器情况下的五项指标均基本优于直接使用该分类器时的情形,且保持了较好的平衡性. 总体来说,HLSCE 模型相比于单独使用基分类器,在总体分类正确率等指标方面均有不同程度的提升,且 HLSCE 模型在以 DT 或 NB 为基分类器时,性能最为优越. 测试阶段,平均而言相比于仅使用单一分类器,HLSCE 模型在 7 个分类器上约提升了 1.48% 的总体分类正确率. 其中,HLSCE 模型在以 NB 为基分类器时,总体分类正确率最高且优于直接使用该分类器时的情形,同时对两类样本的识别能力也更为平衡. 此外,尽管测试阶段各分类器在五项指标方面相对于训练阶段均有不同程度的降低,但 HLSCE 模型的降低程度明显低于仅使用单一分类器时的情形,在稳定性较高之余依然能够保持较好的平衡性.

从表 5 可以看出，在 kaggle 数据集上的训练阶段，HLSCE 模型以 DT 或 NB 为基分类器，相比与仅使用单一分类器或 HLSCE 模型结合其它基分类器时的性能更为优越，同时 HLSCE 模型在与各基分类器结合的情况下均保持了较好的平衡性. 在测试阶段，各模型相比于训练阶段在五个评价指标上均有不同程度的降低，但 HLSCE 模型依然保持了较好的稳定性，当 HLSCE 模型以 DT 或 NB 为基分类器时，模型的总体分类正确率依然最高，这与训练阶段的情形相似，表明模型的稳定性较好. 此外，HLSCE 模型结合各基分类器均能进一步提升模型的分类性能，在五项评价指标方面均优于单独使用该基分类器时的情况，相比于仅使用基分类器，测试阶段 HLSCE 在 7 个分类器上平均约提升了 3.04% 的总体分类正确率，同时也保持了较好的平衡性.

综合表 4 及表 5 发现，两个数据集上的各单一模型在作为基分类器嵌入 HLSCE 模型后，五项性能评价指标均有不同程度的提升. 此外，在两个数据集上的训练阶段与测试阶段，HLSCE 模型以 DT、NB 为基分类器时均达到了较高的总体分类正确率，同时保持了较好的稳定性与平衡性. 以 DA、logistic 与 k-NN 为基分类器时表现适中，以 SVM-liner 与 SVM-gauss 为基分类器时则表现较差. 而在仅使用 7 个单一分类器时，各分类器间性能差异也与 HLSCE 模型下的情形相似，由此认为基分类器的选择是影响 HLSCE 模型性能的重要因素之一.

最后，根据表 6 也发现，即使在同一数据集中，不同基分类器下的较优 k_{pos}、k_{neg} 组合也不尽相同，且同一基分类器在不同数据集上的较优参数组合也存在差异. 同时，HLSCE 模型的性能也较敏感于参数 k_{pos}、k_{neg} 的选择，不同参数下的模型总体分类正确率差异最高超过了 8%. 因此，在缺乏先验信息的条件下，通过诸如 CS 算法的生物启发式算法对参数进行寻优，能够进一步提升模型的性能.

综上，在大样本数据集中，样本的同一属性在不同局部区域内，对该区域中样本分类的贡献程度往往是不同的. 传统基于小样本设计的信用评估模型，由于样本数量较少，难以利用样本在不同局部区域内的可分类特征，只能在均值层面对样本各属性的可分类特征进行提取，因此模型的分类性能有待提升. HLSCE 模型则根据各样本与 V^+、V^- 的相似程度，结合 CS 算法选择较优的 k_{pos}、k_{neg} 参数组合，将样本划入不同子集后分别进行建模，充分利用了样本在不同局部区域中的可分类特征，因而具有更好的分类能力. 最后，基于两个大样本数据集提供的证据，HLSCE 模型在以 DT 或 NB 为基分类器时，性能最为优越.

5 结束语

信用评估是当前信贷机构管理信用风险的重要方法之一. 随着大数据时代的到来，评估可用的样本规模也在不断增加. 首先提出用以描述大样本分布特征的相关属性集、边界向量等若干概念及定义，并证明了其主要性质. 之后基于 PAKDD 数据集与 kaggle 数据集，研究了大样本环境下样本的相似性分布特征，发现在两个数据集中，两类样本与 V^+、V^- 的距离均具有明显的钟形曲线分布特征，但并不服从同一分布，且均非正态分布. 最后，在上述特征启发下设计了面向大样本信用评估的 HLSCE 模型. HLSCE 模型首先根据两类样本与 V^+、V^- 的相似性，结合 CS 算法寻找较优的划分参数，将整体样本集划分为近 V^+ 样本集、近 V^- 样本集与中类样本集三类，进而在三个子集上分别训练模型，最终组合成完整的大样本混合信用评估模型. 相比于现有的只能在均值层面对大样本的可分类特征进行浅层利用的信用评估模型，HLSCE 模型将相似异类样本划入同一训练子集并分别训练基分类器的机制，能够进一步利用相似异类样本间的可分类特征. 实证研究表明，HLSCE 模型在分别以 7 种流行分类器为基分类器时，其性能均优于仅使用单一分类器时的情形，且能保持较好的平衡性与稳定性. 最后，研究也发现 HLSCE 模型的性能在一定程度上也取决于基分类器的选择和子集划分参数 k_{pos}、k_{neg} 的设定.

参 考 文 献:

[1] Lessmann S, Baesens B, Seow H V, et al. Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research [J]. European Journal of Operational Research, 2015, 247(1): 124 – 136.

[2] Orgler Y E. A credit scoring model for commercial loans [J]. Journal of Money, Credit and Banking, 1970, 2(4): 435 – 445.

[3] Steenackers A, Goovaerts M J. A credit scoring model for personal loans [J]. Insurance: Mathematics and Economics, 1989, 8(1): 31 – 34.

[4] Leonard K J. Empirical Bayes analysis of the commercial loan evaluation process [J]. Statistics & Probability Letters, 1993, 18(4): 289 – 296.

[5] 王春峰, 李汶华. 小样本数据信用风险评估研究 [J]. 管理科学学报, 2001, 4(1): 28 – 32.
Wang Chunfeng, Li Wenhua. Credit risk assessment in commercial banks under less samples [J]. Journal of Management Sciences in China, 2001, 4(1): 28 – 32. (in Chinese)

[6] 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.
Zhang Hongxiang, Mao Zhizhong. 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)

[7] Leong C K. Credit risk scoring with Bayesian network models [J]. Computational Economics, 2016, 47(3): 423 – 446.

[8] Zhao Z, Xu S, Kang B H, et al. Investigation and improvement of multi-layer perception neural networks for credit scoring [J]. Expert Systems with Applications, 2015, 42(7): 3508 – 3516.

[9] 于立勇. 商业银行信用风险评估预测模型研究 [J]. 管理科学学报, 2003, 6(5): 46 – 52.
Yu Liyong. Study on credit risk assessing and forecasting model in commercial bank [J]. Journal of Management Sciences in China, 2003, 6(5): 46 – 52. (in Chinese)

[10] Sermpinis G, Stasinakis C, Rosillo R, et al. European exchange trading funds trading with locally weighted support vector regression [J]. European Journal of Operational Research, 2017, 258(1): 372 – 384.

[11] Huang C L, Chen M C, Wang C J. Credit scoring with a data mining approach based on support vector machines [J]. Expert Systems with Applications, 2007, 33(4): 847 – 856.

[12] 庞素琳, 巩吉璋. C5.0 分类算法及在银行个人信用评级中的应用 [J]. 系统工程理论与实践, 2009, 29(12): 94 – 104.
Pang Sulin, Gong Jizhang. C5.0 classification algorithm and its application on individual credit score for banks [J]. Systems Engineering: Theory & Practice, 2009, 29(12): 94 – 104. (in Chinese)

[13] Henley W E. Construction of a k-nearest-neighbour credit-scoring system [J]. IMA Journal of Management Mathematics, 1997, 8(4): 305 – 321.

[14] Kozeny V. Genetic algorithms for credit scoring: Alternative fitness function performance comparison [J]. Expert Systems with Applications, 2015, 42(6): 2998 – 3004.

[15] Hsieh N C. Hybrid mining approach in the design of credit scoring models [J]. Expert Systems with Applications, 2005, 28(4): 655 – 665.

[16] 肖　进, 刘敦虎, 顾　新, 等. 银行客户信用评估动态分类器集成选择模型 [J]. 管理科学学报, 2015, 18(3): 114 – 126.
Xiao Jin, Liu Dunhu, Gu Xin, et al. Dynamic classifier ensemble selection model for bank customer's credit scoring [J]. Journal of Management Sciences in China, 2015, 18(3): 114 – 126. (in Chinese)

[17] Zhou Y. Design dynamic credit risk model with fuzzy rules for auto dealers [J]. International Journal of Business Forecasting and Marketing Intelligence, 2017, 3(3): 248 – 258.

[18] Finlay S. Multiple classifier architectures and their application to credit risk assessment [J]. European Journal of Operational Research, 2011, 210(2): 368 – 378.

[19] Hand D J. Classifier technology and the illusion of progress [J]. Statistical Science, 2006, 21(1): 1 – 14.

[20] Franke B, Plante J F, Roscher R, et al. Statistical inference, learning and models in big data [J]. International Statistical Review, 2016, 84(3): 371 – 389.

[21] 王天思. 大数据中的因果关系及其哲学内涵 [J]. 中国社会科学, 2016, (5): 22 – 42.

Wang Tiansi. Causality in big data and its philosophical connotations [J]. Social Sciences in China, 2016(5): 22 – 42. (in Chinese)
[22] Lehmann E L. Elements of Large-sample Theory [M]. New York: Springer Science & Business Media, 1999.
[23] García V, Marqués A I, Sánchez J S. An insight into the experimental design for credit risk and corporate bankruptcy prediction systems [J]. Journal of Intelligent Information Systems, 2015, 44(1): 159 – 189.
[24] Bellotti T, Crook J. Support vector machines for credit scoring and discovery of significant features [J]. Expert Systems with Applications, 2009, 36(2): 3302 – 3308.
[25] Tsaih R, Liu Y J, Liu W, et al. Credit scoring system for small business loans [J]. Decision Support Systems, 2004, 38(1): 91 – 99.
[26] Harris T. Credit scoring using the clustered support vector machine [J]. Expert Systems with Applications, 2015, 42(2): 741 – 750.
[27] 方匡南，吴见彬，朱建平，等. 信贷信息不对称下的信用卡信用风险研究 [J]. 经济研究，2010，S1: 97 – 107.
Fang Kuangnan, Wu Jianbin, Zhu Jianping, et al. Forecasting of credit card credit risk under asymmetric information based on nonparametric random forests [J]. Economic Research Journal, 2010, S1: 97 – 107. (in Chinese)
[28] Lessmanna S, Seowb H, Baesenscd B, et al. Benchmarking state-of-the-art classification algorithms for credit scoring: A ten-year update [C] //Credit Research Centre, Conference Archive, 2013.
[29] Kar A K. Bio inspired computing: A review of algorithms and scope of applications [J]. Expert Systems with Applications, 2016, 59: 20 – 32.
[30] Yang X S, Deb S. Cuckoo search via Lévy flights [C] // Nature & Biologically Inspired Computing, 2009. NaBIC 2009. World Congress on, 2009, 210 – 214.
[31] Lee T S, Chen I F. A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines [J]. Expert Systems with Applications, 2005, 28(4): 743 – 752.
[32] Chawla N V, Bowyer K W, Hall L O, et al. SMOTE: Synthetic minority over-sampling technique [J]. Journal of Artificial Intelligence Research, 2011, 16(1): 321 – 357.

A hybrid large sample credit evaluation model based on combining similar samples

ZHANG Run-chi, DU Ya-bin, XUE Li-guo, XU Yuan-hao, WU Xin-hong
School of Business, Nanjing University, Nanjing 210093, China

Abstract: Most of the current credit evaluation models designed for large samples have no in-depth consideration of the distribution characteristics of large samples, and just simply apply traditional evaluation methods to large samples. This paper firstly proposes the concept and definition of the related attribute set, boundary vector and so on to describe the distribution characteristics of large samples and proves their main attributes. Then the characteristics of sample distribution are studied in the aspect of similarity based on two large sample data sets. Finally, a hybrid large sample credit evaluation model: HLSCE model is designed. The key idea of HLSCE model is that in large sample data sets, the contribution of the same attribute of samples in different local areas are different to classification performance. Specifically, HLSCE model divides, with biological heuristic algorithm, the whole data set into several subsets according to the similarity between samples and boundary vectors, and then trains the basic classifiers respectively on each subsets. The empirical study shows that compared with the existing representative credit evaluation models, our HLSCE model has a higher classification accuracy, as well as a better balance and stability.

Key words: credit risk; credit evaluation; large sample; boundary vector