银行客户信用评估动态分类器集成选择模型①

肖 进[1,2]，刘敦虎[3]，顾 新[1,4]，汪寿阳[2]
(1. 四川大学商学院，成都 610064; 2. 中国科学院数学与系统科学研究院，北京 100190;
3. 成都信息工程学院管理学院，成都 610225; 4. 四川大学软科学研究所，成都 610064)

摘要: 现实的银行客户信用评估数据常包含大量的缺失值,这在很大程度上影响了信用评估模型的性能.针对已有模型的不足,提出了面向缺失数据的动态分类器集成选择模型 DCESM.该模型充分利用数据集中所包含的已知信息,在训练信用评估模型之前不需要事先对缺失数据进行预处理,从而减少了对数据缺失机制假设以及数据分布模型的依赖.从 UCI 数据库中选择两个银行信用卡业务信用评估数据集进行实证分析,结果表明,与 4 种常用的基于插补法的多分类器集成模型以及 1 种直接面向缺失数据建模的集成模型相比,DCESM 模型能够取得更好的客户信用评估性能.

关键词: 信用评估; 缺失数据; 动态分类器集成选择
中图分类号: F830.91 **文献标识码:** A **文章编号:** 1007-9807(2015)03-0114-13

0 引 言

随着我国逐步从计划经济向社会主义市场经济体制转型,商业银行的信贷业务发展迅速.一方面银行面临的信用欺诈行为越来越多,给整个银行业带来了巨大的损失,据 Visa 和 MasterCard 两大信用卡联盟的统计,全球信用卡欺诈涉及金额已超过 100 亿美元[1].另一方面,许多中小企业急需资金,但由于自身规模较小、经营风险高等因素,在向银行申请贷款时常常遭到拒绝[2].因此,建立科学的客户信用评估模型,不仅能够为银行经理人员提供重要的决策支持,减少银行损失,也能够让那些急需资金、信用较好的企业及时获得贷款.

银行客户信用评估的实质是客户分类问题[3],根据客户违约风险的大小将不同客户分成两类 "信用好"的客户和"信用差"的客户.国内外对银行客户信用评估开展了深入的研究,常用的信用评估模型有人工神经网络(ANN)、贝叶斯网络、判别分析、Logistic 回归、支持向量机(SVM)等.在国外,Sustersic 等[4] 设计了基于 ANN 的模型来解决当金融机构缺乏已有研究中常用的信用评估数据情况下的信用评估问题,Panigrahi 等[5] 将 D-S 证据理论(dempster-shafer theory)和贝叶斯网络相结合,提出了信用评估融合模型,Desai 等[6] 深入比较了 ANN 和 logistic 回归的客户信用评估性能,Chen 和 Li[7] 将 SVM 和 4 种常用的特征选择方法相结合来构建不同版本的信用评估模型.在国内,王春峰等[8] 将判别分析应用于商业银行信用风险评估中,王春峰和万海晖[9]、于立勇[10] 和吴冲等[11] 都建立了基于 ANN 的银行信用风险评估模型,方洪全和曾勇[12] 建立了 4 水平的线性判别和 Logistic 回归并用于银行信用评估,郭英见和吴冲[13] 提出了基于信息融合的商业银

① 收稿日期: 2013-12-15; 修订日期: 2014-11-19.
基金项目: 国家自然科学基金资助项目(71471124; 71101100; 71273036); 四川省社科规划资助项目(SC14C019); 四川省教育厅创新团队资助项目(13TD0040); 四川大学优秀青年基金资助项目(2013SCU04A08); 四川大学学科前沿与交叉创新资助项目(skqy201352).
作者简介: 肖 进(1983—),男,四川广安人,博士后,副教授. Email: xjxiaojin@126.com

行信用风险评估模型,吴冲和夏晗[14]、姚潇和余乐安[15]将 SVM 用于信用风险评估中.

近年来,随着客户信用评估理论和实践研究的深入,人们发现用于信用评估的客户数据常常包含许多缺失数据,且会在很大程度上影响到客户信用评估的效果.这些缺失值可能是结构性缺失(例如,一些问题是需要被调查者根据对前面问题的回答而有选择性地回答)或者随机缺失[16].Hand 和 Henley[16]指出,在他们遇到的一个银行信用评估问题中,数据集中共有 3 883 个样本,其中,只有 66 个样本不包含缺失值,而 25 个特征中只有 5 个不包含缺失值,有 2 个特征包含的缺失值甚至超过 2 000.然而,传统的客户信用评估模型如神经网络、支持向量机等都要求模型的训练集数据必须是完整的[17].为了解决这一问题,常用的策略是先对缺失数据进行预处理,使数据集变得完整,然后再训练客户分类模型,其中数据插值方法应用最为广泛,包括单值插补方法如均值替代、线性回归插补、EM 插补以及多重插补方法等[18].例如,Lessmann 和 Voß[19]提出了一种基于支持向量机的层次参考模型用于信用评价,在建立模型之前采用均值替代来处理缺失数据,而 Barakova 等[20]在对客户信用评估数据分析建模之前,对缺失数据采用多重插补方法.上述研究在构建客户信用评估模型时,都采用单一的分类模型,为了提高模型的性能,近年来,也有部分学者将多分类器集成技术[21]用于客户信用评估中,如 Paleologo 等[22]提出了用于客户信用评估的 Subagging 集成模型,该模型首先使用缺失值所对应特征中观测值的最大值或者最小值对其进行插补,然后构建一个多分类器系统.然而,插补法仍有不足之处,常用的插补方法都是基于随机缺失假设的,且都需要假定数据服从某一分布模型,但在实际应用中,各种缺失方式经常是交织在一起的,采用的假设、模型不合理,将影响后继分类器的学习效果[23].

为了弥补插补法的不足,近年来有学者尝试使用集成学习技术直接构建面向缺失数据的分类模型.如 Krause 和 Polikar[24]提出了直接为含有缺失值的数据集进行分类的集成模型(learn^{++} for missing features,LMF).该模型首先利用随机子空间方法 RSS 来选择一系列特征子集,通过映射计算得到若干训练子集,并在每个训练子集上产生一个弱的分类器构成基分类器池,然后对于每一个待分类样本 x^*(可能含有缺失特征),用基分类器池中没有使用 x^* 中的缺失特征的那些基本分类器为 x^* 进行分类,最后将这些分类结果进行投票得到最终分类结果.实验分析表明,该模型具有较好的分类性能.然而,需要指出的是,该模型只着重研究了测试集中包含缺失的情形,没有考虑训练集中也包含缺失的情形,但在现实的客户分类问题中,训练集和测试集通常都存在数据缺失,同时,Mohammed 等[25]指出,当某一测试样本包含较多缺失特征时,很可能在基分类器池中一个可行的分类器都找不到,这时 LMF 方法将无法对其进行分类.

本文提出了面向缺失数据的客户信用评估动态分类器集成选择模型(dynamic classifier ensemble selection for missing values,DCESM).该方法能够充分利用已知数据集中所包含的信息,在进行客户分类之前不需要事先对缺失数据进行处理,从而减少了对数据缺失机制假设以及数据分布模型的依赖.此外 DCESM 也能处理训练集和测试集中同时存在数据缺失的客户信用评估问题.实证分析表明,该模型的性能优于已有的客户信用评估模型.

1　相关理论介绍

1.1　数据缺失的机制

在缺失数据分析中,缺失机制的重要作用曾被长期忽视,直到 Rubin[26]才给予重视并给出了明确的概念.一般来说,数据缺失可以分为 3 种:随机缺失(missing at random,MAR)、完全随机缺失(missing completely at random,MCAR)和非随机缺失(missing not at random,MNAR).

1.1.1　随机缺失

将数据集分成 X_{obs} 和 X_{mis} 两部分,其中,X_{obs} 代表所有的观测数据,X_{mis} 代表所有的缺失数据.用向量 $\boldsymbol{R}=(R_1, R_2, \cdots, R_n)$ 作为响应变量,当 $R_i=1$ 时,表示变量 X_i 被观测到,$R_i=0$ 时,表示变量 X_i 为缺失数据.若满足

$$P(R \mid X_{\text{obs}}, X_{\text{mis}}) = P(R \mid X_{\text{obs}}, \varphi) \tag{1}$$

式中，φ 是与数据集中任何变量都无关的参数，则称数据集中的数据是随机缺失的. 即数据的缺失是由数据集中与此变量有关的其他变量的取值决定的，与数据本身的取值无关. 随机缺失是比较常见的缺失类型，现有的多种缺失数据处理方法都假定数据是随机缺失的.

1.1.2 完全随机缺失

如果数据集 D 满足

$$P(R \mid X_{obs}, X_{mis}) = P(R/\varphi) \quad (2)$$

则称数据集 D 中的数据是完全随机缺失的. 它是缺失数据问题中最简单的类型. 缺失现象完全是随机发生的，样本中某一变量的缺失与该变量本身的取值以及样本中其他变量的取值完全无关. 在实际应用中符合这种完全随机缺失的情况非常少见.

1.1.3 非随机缺失

如果数据集的缺失情况不满足以上两种缺失方式，则为非随机缺失，这是缺失问题中最难处理的一种. 在这种情况下，缺失不仅和其他变量的取值有关，还和变量本身的取值有关.

识别缺失数据的产生机制是极其重要的. 首先这涉及到代表性问题. 从统计上说，非随机缺失的数据会产生有偏估计，因此不能很好地代表总体. 其次，它决定数据插补方法的选择. 在实际问题中，很难对数据的缺失情况作出准确的判定，各种缺失情况是经常交织在一起发生. 许多方法都是首先假定数据是随机缺失的.

1.2 缺失数据的处理方法

目前，处理缺失数据的主要方法有个案删除法(listwise deletion，LS)、单值插补法(single imputation)和多重插补法(multiple imputation，MI).

1.2.1 个案删除法

个案删除法是最简单的处理缺失数据的方法. 在这种方法中如果任何一个变量含有缺失数据的话，就把相对应的个案从数据集中删除. 如果缺失值所占比例比较小的话，这一方法十分有效. 然而，这种方法却有很大的局限性. 在实际应用中可能所有的样本都含有缺失数据[27]，此时，该方法不能用. 它是以减少样本量来换取信息的完备，会造成资源的大量浪费，丢弃了大量隐藏在这些对象中的信息. 因此，当缺失数据所占比例较大，特别是当缺失数据是非随机分布的时候，这种方法可能导致数据发生偏离，从而得出错误的结论.

1.2.2 单值插补法

1) 均值替换法(mean substitution，MS)

这一方法将变量的属性分为数值型和非数值型来分别进行处理. 如果缺失值是数值型的，就根据该变量在其他所有对象的取值的平均值来填充该缺失的变量值；如果缺失值是非数值型的，就根据统计学中的众数原理，用该变量在其他所有对象的取值次数最多的值来补齐该缺失的变量值. 但这种方法会产生有偏估计，所以并不被推崇. 使用均值替换法插补缺失数据，对该变量的均值估计不会产生影响. 但这种方法是建立在完全随机缺失(MCAR)的假设之上的，而且会造成变量的方差和标准差变小.

2) 期望值最大化(expectation maximization，EM)插补方法

EM 算法[28]是根据所得观测数据，获得对模型参数估计的方法，其核心思想就是根据已有的数据来递归估计似然函数. EM 算法包括两步：E 步指根据 D_{obs} 和第 t 次迭代得到的模型参数 $\theta^{(t)}$ 来预测 $D_{mis}^{(t)}$；M 步是指根据 D_{obs} 和 $D_{mis}^{(t)}$ 来估计 $\theta^{(t+1)}$. 给定模型参数 θ 的初值 $\theta^{(0)}$，重复 E 步和 M 步，直到参数估计收敛为止，收敛时得到的 $D_{mis}^{(t)}$ 可看作插补值. 作为一种迭代方法，其最主要的缺点在于计算成本较大，还可能出现局部收敛和伪收敛的现象. 此外，EM 算法要求提前给出分布形式，这在数据挖掘的大多数应用中都是难以做到的.

3) 回归插补法(regression imputation，RI)

回归插补法首先需要选择若干个预测缺失值的自变量，然后建立回归方程估计缺失值，即用缺失数据的条件期望值对缺失值进行替换. 与前述几种插补方法比较，该方法利用了数据库中尽量多的信息. 但该方法也有诸多弊端：第 1，它虽然是无偏估计，但是却容易忽视随机误差，低估标准差和其他未知性质的测量值；第 2，研究者必须假设

存在缺失值所在的变量与其他变量存在线性关系,很多时候这种关系是不存在的.

4) K-近邻插补法(K-nearest neighbors imputation, KI)

K-近邻插补法是非参数插补方法,常用的是 1-近邻(即最近邻).对于数据集中的任一缺失值,该方法从训练集中找到离缺失值所在样本最近的样本,并使用该样本的值来对缺失值进行插补.这里的距离测度常采用欧氏距离.需要注意的是,在使用该种方法进行插补时,需要对数据集进行标准化处理,以消除各变量的量纲对样本距离的影响.

1.2.3　多重插补法(MI)

单值插补法用某一确定的值代替缺失值,未能反映出缺失值的不确定性.虽然经过插补后获得完整数据集使得不能用的数据处理方法变为可行,但单值插补造成分布扭曲,可能会使得到的结果偏差较大.为改善这一弊端,美国哈佛大学统计学家 Rubin 提出了多重插补方法[18].首先,MI 用一系列可能的值来替换每一个缺失值,以反映被替换的缺失数据的不确定性.然后,用标准的统计分析过程对多次替换后产生的若干个数据集进行分析.最后,把来自于各个数据集的统计结果进行综合,得到总体参数的估计值.然而,MI 的操作比较复杂,工作量大,与单值插补法相比,成本增加许多.

2　面向缺失数据的动态分类器集成选择模型

2.1　基本思想

事实上,在客户信用评估中,除了常包含缺失值以外,往往还面临许多其他数据问题,比如噪声、类别分布的不平衡(即数据集中不同类别的样本数相差很大)等.本研究主要关注客户信用评估中的缺失值问题,提出了面向缺失数据的动态分类器集成选择模型(DCESM).因此,对于客户信用评估数据中可能存在的类别分布不平衡问题,仍需要通过预处理来平衡类别分布,再建立 DCESM 模型.

假设一个客户信用评估问题包含 n 个特征,它的训练集 D_{train} 和测试集 D_{test} 分别包含 m_1 和 m_2 个样本,且都包含一部分缺失值.此外,所有客户样本被分成 L 类(这里分成几类根据不同的信用评估问题来确定,如分成信用好和信用差 2 类,或者好、中、差 3 类等).DCESM 模型的构建主要包含两个阶段:在训练集上训练基本分类器和对测试样本分类.当训练集包含缺失值时,包含缺失值较少的那些特征,其所包含的信息量就大.直观上来讲,包含缺失值越少的特征,其被选择用来训练模型的可能性就越大.因此,在 DCESM 模型的第一阶段,首先根据样本的缺失比例将 D_{train} 分成 3 个子集: D_1, D_2, D_3(划分为 3 个子集的原因在于这 3 个子集分别对应低水平缺失集、中水平缺失集和高水平缺失集),在每个子集中根据每个特征所包含的缺失值个数的不同为其赋予不同的被选择权重,然后选择 T 个不同的特征子集并通过映射得到一系列训练子集,再在每个训练子集中删除带缺失值的样本,并训练一个基本分类模型;最后,这些基本分类器构成了一个基分类器池(BCP).在测试样本分类阶段,对于每一个测试样本 $x_j^* \in D_{\text{test}}(j = 1, 2, \cdots, m_2)$,模型从训练集 D_{train} 中寻找 x_j^* 的 K 个近邻构成它的局部区域 L_a,然后从 BCP 中选择一些在 L_a 中具有最好分类性能的分类器来对 x_j^* 分类,最后通过加权投票的方式得到 x_j^* 的最终分类结果.DCESM 模型的建模过程见图 1.

值得注意的是,在 DCESM 模型中,如果在训练集 D_{train} 中的缺失值比较少,那么大部分训练样本将会被分配到低缺失水平子集 D_1 中,可以使用较多的训练样本训练出具有较好分类性能的基本分类器,从而保证 DCESM 模型的客户信用评估性能;而如果训练集 D_{train} 中的包含的缺失值较多,那么在 3 个子集 D_1, D_2 和 D_3 都包含一定数量的样本,模型仍然能在包含较少缺失值的训练子集 D_1 和 D_2 上得到具有较好分类性能的基分类器来为测试样本 x_j^* 进行分类.因此,DCESM 模型能够在很大程度上弥补 Krause 和 Polikar[24] 提出的

LMF 模型的不足.

接下来,描述 DCESM 模型详细的建模过程.

2.2 训练基本分类器

为了训练基本分类器,DCESM 模型首先根据样本的缺失比例将 D_{train} 分成 3 个子集.对于每一个样本 $x_i \in D_{\text{train}}$,它的缺失比例 α_i 定义如下

$$\alpha_i = \frac{n_{\text{miss}}}{n},\ i = 1,2,\cdots,m_1 \tag{3}$$

式中 n_{miss} 表示在样本 x_i 中缺失的特征个数,n 是样本特征个数,易知 $0 \leqslant \alpha_i \leqslant 1$.

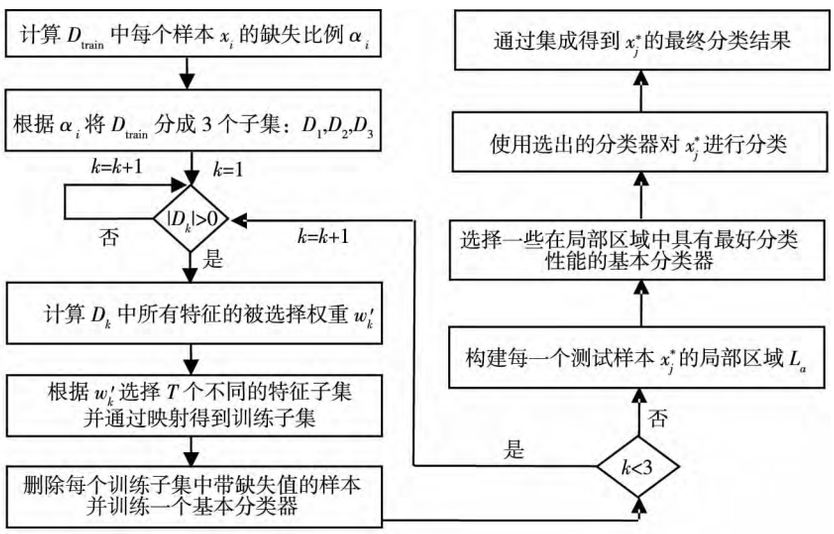

图 1 DCESM 模型的建模过程

Fig. 1 Modeling process of DCESM

在计算出 $\alpha_i(i = 1,2,\cdots,m_1)$ 以后,将所有训练样本按照 α_i 的值从小到大的顺序进行排序.排在越前面的样本,其包含的缺失值越少,甚至不包含缺失.再将 α_i 的取值范围分成 3 个区间:$[0, 1/3)$、$[1/3, 2/3)$、$[2/3, 1]$,并据此将整个训练集 D_{train} 分成 3 个子集 D_1、D_2 和 D_3.因此,D_1 中的样本缺失比例 $\alpha_i \in [0, 1/3)$,类似地,D_3 中的样本缺失比例 $\alpha_i \in [2/3, 1)$.

为了在每个子集 $D_k(k = 1, 2, 3)$ 上训练基本分类器,随机选择一系列特征子集.因为每个特征包含的缺失值往往不同,包含的缺失值越少的特征,其所包含的信息量越大,它被选择的可能性应该越高.对于子集 D_k,如果它是非空的,那么特征 $f_j(j = 1, 2, \cdots, n)$ 的被选择权重 w_{kj} 通过下面公式计算

$$w_{kj} = \frac{|D_k|}{p_{kj}},\ j = 1,2,\cdots,n;\ k = 1,2,3 \tag{4}$$

式中:p_{kj} 是子集 D_k 中特征 f_j 所在列缺失值的个数,$|D_k|$ 为子集 D_k 的样本数.特别地,如果 $p_{kj} = 0$,也就是说,特征 f_j 所在列不包含缺失值,直接令 $w_{kj} = 2$,其目的是为了给它赋予一个比包含缺失值的特征大得多的被选择权重.最后,将权重向量 $\boldsymbol{w}_k = (w_{k1}, w_{k2}, \cdots, w_{kn})$ 进行归一化,特征 f_j 最终的被选择权重为

$$w_{kj}' = \frac{w_{kj}}{\sum w_{kj}},\ j = 1,2,\cdots,n \tag{5}$$

采用遗传算法中的轮盘赌(roulette wheel selection)[29] 选择机制,根据各个特征的被选择权重向量 $\boldsymbol{w}_k' = (w_{k1}', w_{k2}', \cdots, w_{kn}')$,从原始特征空间中选择 T 个不同的特征子集(每个特征子集中的特征数等于原始特征空间特征数的一半,此时,所能产生的候选特征子集的个数是最多的[30]).被选择权重越大的特征,其被抽中的可能性就越大,因此包含该特征的特征子集就越多,它在整个基本分类器训练阶段发挥的作用就越大.再对于每一个选出的特征子集,从原始训练集中将该子集包含的特征所在的那些列提取出来,构成一个训练子集.事实上,此时得到的训练子集中可能还包含一些缺失样本.由于常用的分类模型如神经网络,支持向量机以及 logistic 回归都要求训练集不能包含缺失值.因此,在 DCESM 模型中,将包含缺失的样本从训练子集中删除.如果剩余的训练子集非空,则使用该子集训练一个基本分类器.最后,所有训练出的基本分类器构成一个基本分类

器池(base classifier pool, BCP).值得注意的是,这里的 BCP 中的基分类器个数在不同的数据集上可能不同,这主要取决于数据集中的缺失值的多少以及训练集的样本容量.

2.3　分类测试样本

在本小节中,模型将用训练得到的基分类器为所有测试样本进行分类.本文提出的 DCESM 模型属于动态分类器集成选择策略[3],这一类集成模型的基本思路是从 BCP 中为每一个测试样本 $x_j^* \in D_{\text{test}}(j=1,2,\cdots,m_2)$ 选择一个恰当的分类器子集来为其分类.

为了实现上述过程,模型首先从训练集 D_{train} 中寻找 x_j^* 的 K 个近邻构成它的局部区域 L_a.本文选择欧式距离来测度 x_j^* 与训练集中样本之间的距离.然而,由于 x_j^* 和 D_{train} 中的样本可能包含缺失值,不能直接计算它们之间的欧氏距离.为了解决这一问题,对于每一个测试样本 $x_j^*=(x_{j,1}^*,x_{j,2}^*,\cdots,x_{j,n}^*)$ 和任意训练样本 $x_i=(x_{i,1},x_{i,2},\cdots,x_{i,n})$,本文定义两个缺失值指示向量 $\boldsymbol{u}_j=(u_{j,1},u_{j,2},\cdots,u_{j,n})$ 和 $\boldsymbol{v}_i=(v_{i,1},v_{i,2},\cdots,v_{i,n})$ 如下

$$u_{jk}=\begin{cases}0, & \text{若 } x_{j,k}^* \text{ 缺失}\\ 1, & \text{否则}\end{cases}\quad k=1,2,\cdots,n$$
$$v_{ip}=\begin{cases}0, & \text{若 } x_{i,p} \text{ 缺失}\\ 1, & \text{否则}\end{cases}\quad p=1,2,\cdots,n \tag{6}$$

再假设在测试样本 x_j^* 中只有 s ($s\leqslant n$) 个特征没有缺失,即 $u_{je_1}=u_{je_2}=\cdots=u_{je_s}=1$,这里,$1\leqslant e_t\leqslant n, t=1,2,\cdots,s$. 最后,测试样本 x_j^* 和训练集 D_{train} 中的样本 x_i 间的距离计算如下

$$d_{j,i}=\begin{cases}\sqrt{\sum_{t=1}^{s}(x_{j,e_t}^*-x_{i,e_t})^2}, \\ \text{若 } v_{i,e_1}=v_{i,e_2}=\cdots=v_{i,e_s}=1\\ \infty, \quad \text{否则}\end{cases} \tag{7}$$

式(7)表明,当两个样本间的距离无法计算时,令其距离为∞.

通过式(7)找到 x_j^* 的局部区域 L_a 后,模型使用训练得到的所有基分类器对 L_a 中的样本进行分类并计算每个基分类器的分类精度,选择 N 个($N=|\text{BCP}|/2$,其中 $|\text{BCP}|$ 为基分类器池 BCP 中的分类器个数)具有较高分类精度的分类器.假设选出的 N 个基分类器在 L_a 上的分类精度分别为 $acc_1, acc_2, \cdots, acc_N$. 最后,使用 N 个基分类器对测试样本 x_j^* 进行分类.对于每一个基分类器 $C_r(r=1,2,\cdots,N)$,它会输出一个概率估计 $P(f(x_j^*)=c\mid C_r)$,表示 x_j^* 属于第 c ($c=1,2,\cdots,L$) 类的概率值.而样本 x_j^* 的最终分类结果可以通过加权投票法得到

$$\rho=\arg_c\left(\max_{c=1,2,\cdots,L}\left(\sum_{r=1}^{N}z_r P(f(x_j^*)=c\mid C_r)\right)\right) \tag{8}$$

式中,$z_r=\dfrac{acc_r}{\sum acc_r}$, $r=1,2,\cdots,N$. 也就是说,首先计算 N 个基本分类器对测试样本 x_j^* 属于每一类的概率预测值的加权算术平均,然后找到平均值最大的那一个类(假设为第 l 类),最后将 l 作为 x_j^* 类别的最终集成预测结果.

2.4　模型详细建模步骤

DCESM 模型的详细建模步骤如下:

步骤 1　用式(3)计算训练集 D_{train} 中每个样本的缺失比例 α_i,根据 α_i 的大小将整个训练集分成 3 个子集 D_1、D_2 和 D_3;

步骤 2　对于每一个子集 D_k($k=1,2,3$),如果 D_k 非空,那么

1) 根据式(5)计算所有特征的被选择权重向量 $\boldsymbol{w}_k'=(w_{k1}',w_{k2}',\cdots,w_{kn}')$;

2) 根据 $\boldsymbol{w}_k'$ 从原始特征空间抽取 T 个不同的特征子集,进而通过映射得到 T 个训练子集 $S_i(i=1,2,\cdots,T)$;

3) 删除 S_i 中带缺失的样本,如果 S_i 非空,训练一个基分类器加到基分类器池 BCP 中;

步骤 3　对于每一个测试样本 $x_j^*\in D_{\text{test}}$,

1) 根据式(7)从整个训练集 D_{train} 中找到 x_j^* 的 K 个近邻构成其局部区域 L_a;

2) 使用 BCP 中所有基分类器为 L_a 中的样本进行分类并选择一半具有最高分类精度的基分类器;

3) 使用所选的基分类器对 x_j^* 进行分类,进而根据式(8)得到 x_j^* 的最终分类结果.

3　实证分析

为了分析本文提出的 DCESM 模型的客户信

用评估性能，选择 UCI 数据库[32] 中的 2 个信用评估数据集进行实证分析. 同时，将 DCESM 模型的客户信用评估性能与 4 种基于插补法的集成模型进行比较，这 4 种模型首先分别使用均值插补(MS)、EM 插补(EM)、回归插补(RI) 和 K-近邻插补(KI) 来对缺失值进行插补，然后使用 Subagging 方法[22] 构建多分类器集成模型. 上述 4 种方法，依次简记为 MS-Subagg，EM-Subagg，RI-Subagg 和 KI-Subagg. 值得注意的是，由于直接删除法具有明显的不足，本文没有将其作为比较对象. 同时，多重插补法因为计算复杂度很高，因此也没有选作比较对象. 最后，本文还将 DCESM 模型的客户信用评估性能与 Krause 和 Polikar[24] 提出的直接面向缺失数据建模的多分类器集成模型 LMF 进行了比较.

3.1 数据集描述

German 数据集是 UCI 数据库中著名的信用评估数据集，是关于德国某银行信用卡业务的数据. 该数据集包含 20 个属性变量和 1 个类别变量 C. 其中，有 7 个数值型属性变量，13 个定性变量，类变量有两个不同的状态{good，bad}，相应地将全部客户划分成两类: 信用好的客户和信用差的客户. 该数据集共有 1 000 个客户样本，其中信用好的客户有 700 个，信用差的客户有 300 个，两类客户比例为 2.33∶1，属于类别不平衡数据集. 该数据集并不包含缺失.

Australia 数据集是 UCI 数据库中另一个著名的信用评估数据集，是关于澳大利亚某银行信用卡业务的数据. 该数据集共有 690 个客户样本，包含两类{good，bad}，即信用好和信用差的客户，其中，信用好的客户样本 383 个，信用差的客户样本 307 个，两者的比例为 1.25∶1，仍然属于类别分布不平衡的数据集. 为了保护商业机密，公开的数据集中对属性名和属性值做了符号代换，共有 14 个属性，其中，定量属性 6 个，定性属性 8 个. 同时，在该数据集中，有 37 个客户样本包含 1 个或者多个缺失值，包含缺失的样本占全部样本的 5.36%.

3.2 实验设置

对于两个数据集中的定性属性，若某属性有 w 个不同的取值状态，则依次令它们取“1”，“2”，…，“w”. 而对于连续属性，本文采用 Fayyad 和 Irani[33] 提出的方法对其进行离散化处理.

在数据集 German 中，由于它本身不包含缺失值，因此，采用人工机制来产生缺失数据. 前面已经指出，在 3 种缺失机制中，随机缺失是比较常见的，且现有的很多缺失数据处理方法都假设数据是随机缺失的. 因此在这里，采用随机缺失机制，令缺失比例 β 分别取 5%，10%，20% 和 30%，然后按下面过程产生缺失数据: 对于每个样本，产生随机数 r，若 $r < \beta$ 则从样本中随机选择两个特征变量 x_i，x_j，$i < j$，若 $x_i = 1$，则令 x_j 为缺失. 由于 x_j 的缺失只与变量 x_i 的取值有关，而与其自身的取值无关. 因此，通过这种方法可以得到一个随机缺失数据集.

本文选择支持向量机作为基本的分类算法. 对于支持向量机，最重要的就是核函数的选择[34]. 通过实验比较发现，径向基核函数能够取得最好的效果，因此选取径向基函数形式的核函数. 在 DCESM 模型中有两个重要参数: 构成局部区域的近邻个数 K 和每个子集 $D_k(k = 1, 2, 3)$ 上选择的不同的特征子集的个数 T. 对这两个参数进行敏感性分析，让 K 取 7 种不同的值: 3，5，7，9，11，13 和 15，让 T 也取 7 种不同的值: 10，15，20，25，30，35 和 40，并分别在两个数据集上进行实验. 结果发现，随着 K 的变化，DCESM 模型的客户信用评估性能表现出一定的水平波动，当 $K = 5$ 时，表现出最好的性能; 随着 T 值的增加，DCESM 模型的性能先随之提高，当 T 达到 30 左右时，其性能最好，而当 T 进一步增加时，DCESM 模型的性能则相对比较稳定. 因此，取 $K = 5$，$T = 30$. 进一步地，对于其他 4 种基于插补法的分类器集成模型以及 1 种直接面向缺失数据建模的 LMF 模型，Paleologo 等[22] 发现对于 Subagging 集成模型，基分类器的个数介于 20 — 50 能取得比较好的分类性能，因此，令这 5 种模型的基分类器数取最大值 50.

此外，两个数据集的类别分布都是不平衡的，如果直接用这样的数据来构建客户信用评估模型，那么得到的模型可能会倾向于将全部客户样本归为多数类，即信用好的类. 用于处理类别分布不平衡的技术有很多，包括随机向上抽样、随机向下抽样以及混合抽样等[35]. 由于 Subagging 集成模型已经在构建训练子集的过程中，考虑了类别不平衡问题，因此，在构建 4 种基于插补法的分类

器集成模型 MS-Subagg,EM-Subagg,RI-Subagg 和 KI-Subagg 时不需要再对数据集进行处理.而对于 LMF 模型以及本文提出的 DCESM 模型,因为它们在建模过程中没有考虑类别分布的不平衡问题,因此,为了公平起见,采用随机向上抽样来平衡类别分布,再构建模型.此外,为了比较 DCESM 模型与另外 5 种模型的客户信用评估性能,采用 10 层交叉验证技术,该方法将整个数据集随机分成 10 等份,依次取其中 1 份作为测试集,而余下的 9 份作为模型训练集,如此循环 10 次,称之为一次 10 层交叉验证.最后,所有实验均是在 matlab7.0 软件平台上实现,同时,每一个分类结果均是取 10 次 10 层交叉验证的平均值.

3.3　模型评价准则

现实的客户信用评估数据往往是类别分布不平衡的,在这样的情况下,仅仅使用传统的总体分类精度指标来评价模型的信用评估性能是远远不够的.为了更好地评价各种模型的性能,引入如表 1 所示的评价矩阵,在此基础上,采用 4 个常用的评价准则[8]

1) 总的准确率 = $(TP + TN)/(TP + FN + FP + TN) \times 100\%$;

2) ROC 曲线及 AUC 值.针对两类问题的 ROC 曲线是一个真正率 — 伪正率图,其中横轴表示伪正率 = $FP/(FP + TN) \times 100\%$,纵轴表示真正率 = $TP/(TP + FN) \times 100\%$. 由于直接比较不同分类模型的 ROC 曲线很不方便,因此,很多时候人们使用 ROC 曲线下方的面积,即 AUC(area under the ROC curve) 值来比较模型的分类性能.

表 1　信用评估效果评价矩阵

Table 1 Evaluation matrix for credit scoring

预测类别	正类	负类	总计
正类(信用差的类)	TP	FN	$TP + FN$
负类(信用好的类)	FP	TN	$FP + TN$
总计	$TP+FP$	$FN+TN$	$TP+FN+FP+TN$

注: TP 代表实际为正类,预测为正类的样本个数; FN 代表实际为正类,预测为负类的样本个数; FP 代表实际为负类,预测为正类的样本个数; TN 代表实际为负类,预测为负类的样本个数.

3) 正类(信用差的类) 准确率 = $TP/(TP + FN) \times 100\%$;

4) 负类(信用好的类) 准确率 = $TN/(FP + TN) \times 100\%$.

3.4　结果分析

表 2 给出了 6 种模型在 German 数据集上信用评估性能. 从表中可以看出: 1) 当缺失水平 $\beta = 5\%$,DCESM 模型在 4 个评价准则上都优于其他 5 种模型; 2) 当缺失水平 $\beta = 10\%$,DCESM 在总的准确率、AUC 以及负类准确率等 3 个评价准则上优于其他 5 种模型,只有在正类准确率上比模型 RI-Subagg 要差; 3) 当缺失水平 $\beta = 20\%$,DCESM 模型在所有评价准则上优于另外 5 种多分类器集成模型; 4) 当缺失水平 $\beta = 30\%$,DCESM 在 AUC 和正类准确率上优于其他模型,而它的总的准确率和负类准确率要比模型 EM-Subagg 差. 在银行客户信用评估中,与负类准确率相比,往往最关心正类准确率,即信用差的客户的分类准确率,这样才能最大限度的减少银行的损失.因此,根据上面的分析可知,DCESM 模型在 German 数据集上的整体客户信用评估性能要优于 4 种基于插补法的集成模型 KI-Subagg, MS-Subagg, EM-Subagg 和 RI-Subagg,也优于直接面向缺失数据建模的 LMF 模型.

此外,表 2 还给出了 4 种不同的缺失水平下,6 种模型在每个评价准则下的性能排序,而表 2 的最后一行给出了 6 种模型各自的平均排序,这也可以作为评价各个模型整体信用评估性能的一个标准,平均排序数越小,表示模型的排序越靠前,它的信用评估性能就越好.因此,据此将 6 种模型在 German 数据集上的信用评估性能从高到低排序: DCESM, EM-Subagg, RI-Subagg, MS-Subagg, LMF 和 KI-Subagg.值得注意的是,EM-Subagg 模型的性能只比 DCESM 模型要差,它在 4 种基于插补法的多分类器集成模型中, 性能是最好的, 而 LMF 和 KI-Subagg 模型的性能是最差的.

图 2 给出了 6 种模型在 German 数据集上的信用评估性能变化趋势.从中可以看出,随着数据缺失比例的增加,6 种模型的性能都跟着下降.特别地,当缺失比例较低时(比如 $\beta = 5\%$),直接面向缺失数据建模的多分类器集成模型 LMF 也具有较好的性能, 它的整体性能仅次于 DCESM 和 EM-Subagg 模型.然而,随着缺失比例的增加,当

$\beta \geqslant 30\%$ 时，LMF 模型的性能变得很差.DCESM 与 LMF 模型都直接面向缺失数据构建集成模型，这一结果也证明了 DCESM 模型能够在很大程度上克服 LMF 模型的不足，在数据集中的缺失水平较高的情况下，能够取得比 LMF 模型更好的客户信用评估性能.

表 2 German 数据集上 6 种模型的客户信用评估性能比较

Table 2 Performance comparison of six credit scoring models in German dataset

缺失比例 β(%)	评价准则	LMF	DCESM	KI-Subagg	MS-Subagg	EM-Subagg	RI-Subagg
5	总的准确率	0.720 0(3)	**0.736 7**(1)	0.703 3(6)	0.716 7(4)	0.722 0(2)	0.714 8(5)
	AUC	0.872 4(3)	**0.878 9**(1)	0.844 7(6)	0.866 2(4)	0.873 2(2)	0.865 3(5)
	正类准确率	0.616 7(3)	**0.641 1**(1)	0.597 9(6)	0.611 7(4)	0.598 0(5)	0.620 9(2)
	负类准确率	0.764 3(3)	**0.777 7**(1)	0.748 5(6)	0.761 7(4)	0.775 2(2)	0.755 0(5)
10	总的准确率	0.673 0(6)	**0.720 5**(1)	0.677 7(5)	0.699 0(4)	0.700 0(3)	0.710 7(2)
	AUC	0.830 0(5)	**0.866 1**(1)	0.805 4(6)	0.831 8(4)	0.833 4(3)	0.859 2(2)
	正类准确率	0.583 3(4)	0.601 2(2)	0.561 5(6)	0.593 8(3)	0.570 5(5)	**0.631 0(1)**
	负类准确率	0.711 4(6)	**0.760 6**(1)	0.727 5(5)	0.744 1(4)	0.755 5(2)	0.744 8(3)
20	总的准确率	0.658 0(5)	**0.705 7**(1)	0.658 3(4)	0.655 3(6)	0.675 0(3)	0.685 0(2)
	AUC	0.821 5(6)	**0.853 3**(1)	0.823 0(5)	0.839 0(3)	0.848 4(2)	0.831 7(4)
	正类准确率	0.538 0(6)	**0.607 0**(1)	0.544 6(5)	0.553 4(3)	0.545 9(4)	0.568 3(2)
	负类准确率	0.709 4(4)	**0.748 0**(1)	0.707 1(5)	0.699 0(6)	0.730 3(3)	0.735 0(2)
30	总的准确率	0.633 1(6)	0.675 9(2)	0.646 7(5)	0.653 1(4)	**0.677 2**(1)	0.660 0(3)
	AUC	0.799 0(6)	**0.835 0**(1)	0.807 1(5)	0.815 9(4)	0.823 3(2)	0.821 3(3)
	正类准确率	0.506 3(6)	**0.582 6**(1)	0.521 7(5)	0.546 7(3)	0.544 7(4)	0.554 2(2)
	负类准确率	0.687 4(6)	0.715 9(2)	0.700 2(4)	0.698 7(5)	**0.734 0**(1)	0.705 3(3)
平均排序		5.125	1.188	5.250	4.063	2.750	2.875

注：表中黑体数字表示每一行的最大值，括号中的数值是每一行中 6 种模型在相应评价准则上的排序值.

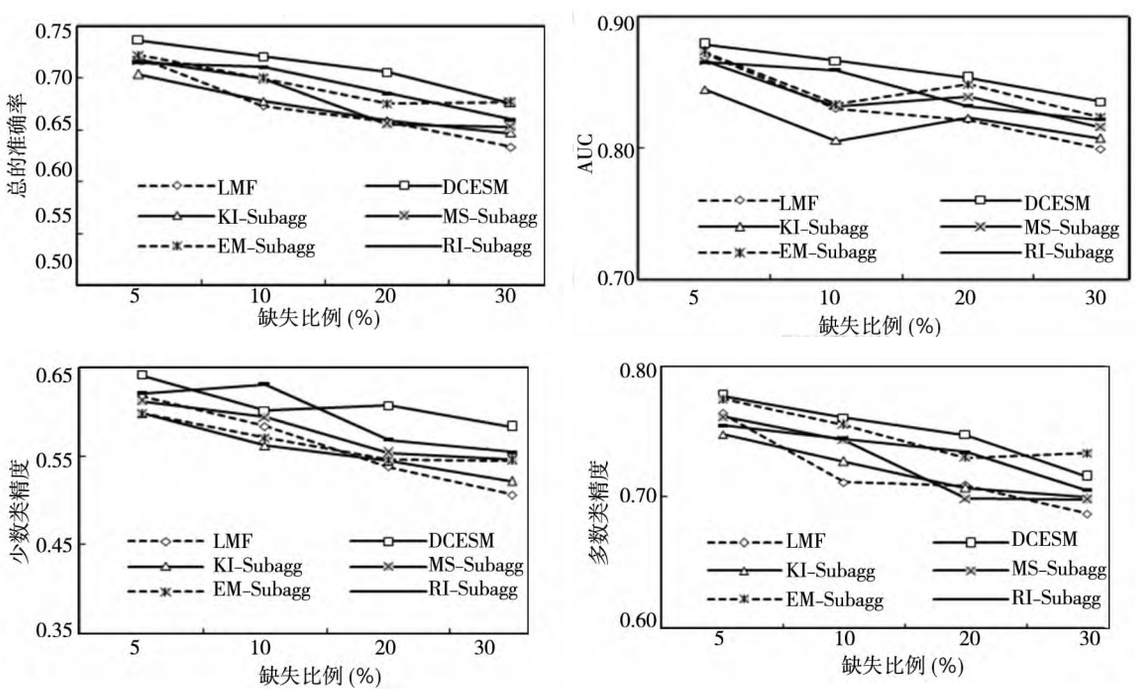

图 2 6 种模型的客户信用评估性能变化趋势

Fig. 2 Trend of credit scoring performance for six models

再则,图 3 给出了 6 种模型在 Australia 数据集上的 ROC 曲线.其中,ROC 曲线位置越高,与 X-轴围成的面积越大,代表对应的模型信用评估性能越好.从图 3 可以看出,6 种信用评估模型的 ROC 曲线之间存在一定程度的交叉,但总的来说,在绝大多数时候 DCESM 的 ROC 曲线位于最上方,其次是 EM-Subagg 模型和 LMF 模型的 ROC 曲线,处于最下面的是 MS-Subagg 模型的 ROC 曲线.因此,可以大致判断,在 Australia 数据集上,DCESM 模型具有最好的信用评估性能,而 MS-Subagg 模型的信用评估性能是最差的.

最后,表 3 给出了 6 种客户信用评估模型在 Australia 数据集上的总的准确率、AUC 值、正类准确率以及负类准确率.表中的黑体数字表示每一列对应的最大值.从表中可以看出,DCESM 模型在总的准确率、AUC 值以及正类准确率(即信用差的客户准确率) 3 个评价准则上要优于另外的 5 种模型,而只有它的正类准确率要低于 LMF 模型.由此得出结论,在 Australia 数据集上,本文提出的 DCESM 模型也能取得最好的信用评估性能.

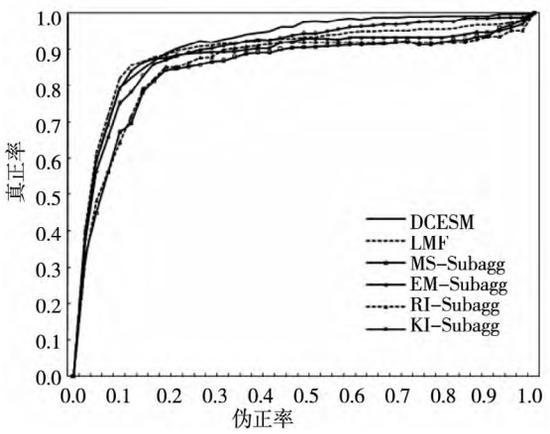

图 3　6 种模型在 Australia 数据集上的 ROC 曲线

Fig. 3　ROC curves of six models in Australia dataset

表 3　6 种模型在 Australia 数据集上信用评估性能比较

Table 3 Performance comparison of six credit scoring models in Australia dataset

模型	总的准确率	AUC	正类准确率	负类准确率
DCESM	**0.887 8**	**0.942 3**	**0.876 2**	0.896 7
LMF	0.880 2	0.913 8	0.857 0	**0.898 0**
KI-Subagg	0.867 5	0.901 5	0.835 2	0.892 3
MS-Subagg	0.857 3	0.889 0	0.826 0	0.881 4
EM-Subagg	0.870 7	0.923 4	0.863 2	0.876 5
RI-Subagg	0.863 2	0.894 2	0.826 3	0.891 6

4　结束语

建立科学的客户信用评估模型,能够为银行经理人员提供重要的决策支持,减少银行损失,意义十分重大.本文主要关注缺失数据环境下的银行客户信用评估建模.针对已有的基于插补法的以及直接面向缺失数据建模的信用评估模型不足,提出了面向缺失数据的动态分类器集成选择模型 DCESM.从 UCI 数据库中选择两个银行信用卡业务信用评估数据集进行试验分析,结果发现,不管是在人工产生缺失数据的 German 数据集还是在本身包含缺失值的 Australia 数据集上,DCESM 模型的客户信用评估性能都优于常用的 4 种基于插补法的多分类器集成模型 KI-Subagg, MS-Subagg,EM-Subagg 和 RI-Subagg,也优于直接面向缺失数据建模的多分类器集成模型 LMF,从而为银行客户信用评估建模提供了一种新的思路.

参考文献:

[1] 陈　雷. 国际信用卡欺诈与预防 [J]. 中国信用卡, 2004, (6): 43-47.

Chen Lei. International credit card fraud and prevention [J]. China Credit Card, 2004, (6): 43-47. (in Chinese)

[2] 熊 熊，姚传伟，张永杰. 中小企业联合担保贷款的计算实验金融分析 [J]. 管理科学学报，2013，16(3)：88-94.
Xiong Xiong, Yao Chuanwei, Zhang Yongjie. Analysis of scale in SME joint guarantee loans using agent-based computational experiment finance [J]. Journal of Management Sciences in China, 2013, 16(3): 88-94. (in Chinese)

[3] 熊海涛，吴俊杰，刘洪甫，等. 分类中的类重叠问题及其处理方法研究 [J]. 管理科学学报，2013，16(4)：8-21.
Xiong Haitao, Wu Junjie, Liu Hongfu, et al. Towards classificaiton with class overlapping [J]. Journal of Management Sciences in China, 2013, 16(4): 8-21. (in Chinese)

[4] Sustersic M, Mramor D, Zupan J. Consumer credit scoring models with limited data [J]. Expert Systems with Applications, 2009, 36(3): 4736-4744.

[5] Panigrahi S, Kundu A, Sural S, et al. Credit card fraud detection: A fusion approach using Dempster-Shafer theory and Bayesian learning [J]. Information Fusion, 2009, 10(4): 354-363.

[6] Desai V S, Crook J N, Overstreet G A. A comparison of neural networks and linear scoring models in the credit union environment [J]. European Journal of Operational Research, 1996, 95(1): 24-37.

[7] 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.

[8] 王春峰，万海晖，张 维. 商业银行信用风险评估及其实证研究 [J]. 管理科学学报，1998，1(1)：68-72.
Wang Chunfeng, Wan Haihui, Zhang Wei. Credit risk assessment in commercial banks and its test [J]. Journal of Management Sciences in China, 1998, 1(1): 68-72. (in Chinese)

[9] 王春峰，万海晖. 基于神经网络技术的商业银行信用风险评估 [J]. 系统工程理论与实践，1999，19(9)：24-32.
Wang Chunfeng, Wan Haihui. Credit risk assessment in commercial banks using neural networks [J]. System Engineering-Theory & Practice, 1999, 19(9): 24-32. (in Chinese)

[10] 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)

[11] 吴 冲，吕静杰，潘启树，等. 基于模糊神经网络的商业银行信用风险评估模型研究 [J]. 系统工程理论与实践，2005，24(11)：1-8.
Wu Chong, Lü Jingjie, Pan Qishu, et al. Study on credit risk assessment model of commercial banks based on fuzzy neural network [J]. System Engineering-Theory & Practice, 2005, 24(11): 1-8. (in Chinese)

[12] 方洪全，曾 勇. 银行信用风险评估方法实证研究及比较分析 [J]. 金融研究，2004，(1)：62-69.
Fang Hongquan, Zeng Yong. An empirical study and comparative analysis on credit risk evaluation of enterprise [J]. Journal of Financial Research, 2004, (1): 62-69. (in Chinese)

[13] 郭英见，吴 冲. 基于信息融合的商业银行信用风险评估模型研究 [J]. 金融研究，2009，(1)：95-106.
Guo Yingjian, Wu Chong. Credit risk evaluation model for commercial bank based on information fusion [J]. Journal of Financial Research, 2009, (1): 95-106. (in Chinese)

[14] 吴 冲，夏 晗. 基于五级分类支持向量机集成的商业银行信用风险评估模型研究 [J]. 预测，2009，28(4)：57-61.
Wu Chong, Xia Han. Credit risk assessment in commercial banks on five-class support vectormachines ensemble [J]. Forecasting, 2009, 28(4): 57-61. (in Chinese)

[15] 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.
Yao Xiao, Yu Le'an. A fuzzy proximal support vector machine model and its application to credit risk analysis [J]. System Engineering-Theory & Practice, 2012, 32(3): 549-554. (in Chinese)

[16] 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.

[17] Kim J, Hwang K J, Bae J K. Prediction of personal credit rates with incomplete data sets using cognitive mapping [C] // IEEE Computer Society Washington, 2007: 1912-1917.

[18] Rubin D B. Multiple Imputations for Nonresponse in Surveys [M]. New York: John Wiley and Sons, 1987.

[19] Lessmann S, Voß S. A reference model for customer-centric data mining with support vector machines [J]. European Journal of Operational Research, 2009, 199(2): 520-530.

[20] Barakova I, Bostic R W, Calem P S, et al. Does credit quality matter for homeownership? [J]. Journal of Housing Economics, 2003, 12(4): 318-336.

[21] Hansen L K, Salamon P. Neural network ensembles [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1990, 12(10): 993-1001.

[22] Paleologo G, Elisseeff A, Antonini G. Subagging for credit scoring models [J]. European Journal of Operational Research, 2010, 201(2): 490-499.

[23] Jiang K, Chen H, Yuan S. Classification for incomplete data using classifier ensembles [C] // International Conference on Neural Networks and Brain, ICNN&B '05, 2005: 559-563.

[24] Krause S, Polikar R. An ensemble of classifiers approach for the missing feature problem [C] // Int. Joint Conf. on Neural Networks, Port-land, OR, 2003: 553-558.

[25] Mohammed H S, Stepenosky N, Polikar R. An ensemble technique to handle missing data from sensors [C] // IEEE Sensors Applications Symposium Houston, Texas USA, 2006: 101-105.

[26] Rubin D B. Inference and missing data [J]. Biometrika, 1976, 63(3): 581-592.

[27] Lakshminarayan K, Harp S A, Samad T. Imputation of missing data in industrial databases [J]. Applied Intelligence, 1999, 11(3): 259-275.

[28] Dempster A P, Laird N M, Rubin D B. Maximum likelihood from incomplete data via the EM algorithm [J]. Journal of the Royal Statistical Society, 1977, 39(1): 1-38.

[29] **夏桂梅，曾建潮. 一种基于轮盘赌选择遗传算法的随机微粒群算法 [J]. 计算机工程与科学**, 2007, 29(6): 51-54.
Xia Guimei, Zeng Jianchao. A stochastic particle swarm optimization algorithm based on the genetic algorithm of roulette wheel selection [J]. Computer Engineering & Science, 2007, 29(6): 51-54. (in Chinese)

[30] Ho T K. The random space method for constructing decision forests [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998, 20(8): 832-844.

[31] Ko A H R, Sabourin R, Jr Britto A S. From dynamic classifier selection to dynamic ensemble selection [J]. Pattern Recognition, 2008, 41(5): 1718-1731.

[32] Merz C, Murphy P. UCI repository of machine learning databases [EB/OL]. http://www.ics.uci.edu/~mlearn/ MLRepository.html, 1995.

[33] Fayyad U M, Irani K B. Multi-interval discretization of continuous-valued attributes for classification learning [C] // Proceedings of 13th International Joint Conference on Artificial Intelligence, 1993: 1022-1027.

[34] **夏国恩，金炜东. 基于支持向量机的客户流失预测模型 [J]. 系统工程理论与实践**, 2008, 28(1): 71-77.
Xia Guo'en, Jin Weidong. Model of customer churn prediction on support vector machine [J]. System Engineering-Theory & Practice, 2008, 28(1): 71-77. (in Chinese)

[35] He H, Garcia E A. Learning from imbalanced data [J]. IEEE Transactions on Knowledge and Data Engineering, 2009, 21(9): 1263-1284.

[36] Burez J, Van den Poel D. Handling class imbalance in customer churn prediction [J]. Expert Systems with Applications, 2009, 36(3): 4626-4636.

Dynamic classifier ensemble selection model for bank customer's credit scoring

XIAO Jin[1,2], *LIU Dun-hu*[3], *GU Xin*[1,4], *WANG Shou-yang*[2]

1. Business School of Sichuan University, Chengdu 610064, China;
2. Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China;
3. Management Faculty, Chengdu University of Information Technology, Chengdu 610225, China;
4. Soft Science Institute of Sichuan University, Chengdu, 610064, China

Abstract: The data in the bank customer's credit scoring often include lots of missing values, which affect the modeling performance to a large extent. To overcome the deficiencies of existing models, this paper proposes a dynamic classifier ensemble selection model for missing values (DCESM). The model can make full use of the information included in the dataset and does not need to pre-process the missing values before training the model, which decreases the dependence on the hypothesis for data missing mechanism and distribution model. Two credit scoring datasets on bank credit card business from UCI database were selected for our empirical analysis. The results show that the DCESM model outperformed four imputation-based multi-classifiers ensemble models and one ensemble model for missing values.

Key words: credit scoring; missing values; dynamic classifier ensemble selection