基于多分类器动态选择与成本敏感优化集成的电信客户流失预测研究

罗 彬[1] 邵培基[1] 夏国恩[2]
(1. 电子科技大学经济与管理学院; 2. 广西财经学院工商管理系)

摘要: 针对不同样本在特征空间中具有不同的区域特性和不同分类算法之间的预测互补性,在电信客户流失预测理论基础上,融合多分类器动态集成理论和成本敏感学习理论,建立了电信客户流失多分类器集成预测的利润函数,并提出了一类新的基于多分类器动态选择与成本敏感优化集成的电信客户流失预测模型。首先使用 K 均值聚类法聚类训练样本成多个分区;接着使用 NaiveBayes 算法、多层感知机算法和 J48 算法在各分区样本上构建客户流失预测子分类器;最后使用改进人工鱼群算法分别对各分区的子分类器进行成本敏感优化集成。实验结果表明,所提出的基于多分类器动态选择与成本敏感优化集成模型的分类性能不仅优于由训练集全体样本所构建的 3 个单模型,也优于基于改进人工鱼群算法优化集成这 3 个单模型而得到的集成模型。

关键词: 客户流失预测; 多分类器动态选择; 成本敏感优化集成; 成本敏感学习; 人工鱼群算法

中图分类号: C93;TP3 **文献标识码**: A **文章编号**: 1672-884X(2012)09-1373-09

A Study on Prediction of Telecom Customer Churn Based on Dynamic Selection, Optimization and Integration of Cost Sensitivity

LUO Bin[1] SHAO Peiji[1] XIA Guoen[2]
(1. University of Electronic Science and Technology of China, Chengdu, China;
2. Guangxi University of Finance and Economics, Nanning, China)

Abstract: On account that the different samples have the prediction complementarities between different section characters and different classification algorithms in feature space and based on the theory of Telecom customer churn prediction, this paper established the profits functions to predict Telecom customer churn integrating multi-classifiers, and a new customer churn prediction model is put forward in Telecom based on the dynamic selection and optimizing integrating of cost sensitivity. Firstly, the training set samples are clustered into multiple subareas by using K-means clustering algorithm. Then, the customer churn prediction sub-classifiers are established based on the samples in the subareas by using NaiveBayes Algorithm, Multilayer Perceptron and J48 Algorithm, respectively. Finally, the subarea sub-classifiers are integrated and optimized by use of the Improved Artificial Fish-school Algorithm(IAFSA). The experiment results show that the classifying performance of the model based on the dynamic integration of multi-classifiers and optimizing integrating of cost sensitivity not only excels the three single model constructed based on the whole samples, but also excels the model integrating of the three single model by IAFSA.

Key words: customer churn prediction; dynamic selection of multiple classifiers; optimizing integration of cost sensitivity; cost sensitivity learning; artificial fish-school algorithm (AFSA)

收稿日期: 2010-01-29
基金项目: 国家自然科学基金资助项目(70801021);中国博士后科学基金资助项目(20080431276);教育部人文社会科学资助项目(08JC630019)

电信客户流失预测具有 2 个特点：①成本敏感性。若将一个电信流失客户错误预测成非流失客户，进而让客户挽留部门错失挽留机会而造成的损失，远远大于把一个非流失客户错误预测成流失客户所造成挽留资源的浪费[1~3]。由此，这种成本敏感特点决定电信客户流失预测不适合使用基于预测精度的预测方法，因为该方法只适用于不同类别的错误预测损失是相等的情况。②非对称性。电信客户流失预测是属于类别严重不对称的分类问题，即在样本数据中，流失客户数量远远少于不流失客户数量[1~3]，这种数据特征意味着使用传统的基于类别对称假设的分类算法是难以提高稀有类别的预测精度的。针对上述情况，通过文献分析得到，对解决成本敏感性问题和类别不对称性问题最有效手段是使用成本敏感学习理论和模型集成理论。

目前，基于成本敏感学习理论的客户流失预测已引起了学者的研究兴趣。如钱苏丽等[1]和蒋国瑞等[2]都使用了成本敏感学习理论来改进支持向量机算法，提出的基于改进支持向量机的电信客户流失预测模型，都获得了较好的预测性能；XIE 等[3]使用成本敏感学习理论改进了随机森林分类算法，提出的基于改进随机森林算法的银行客户流失预测模型也获得较好效果。与此同时，基于模型集成的客户流失预测研究也受到极大关注。如王纯麟等[4]针对单分类器模型不足，提出一种基于 AdaBoost 组合分类器的电信客户流失预测模型，并取得了较好结果。但该集成模型只采用了 C4.5 作为基分类器算法，可能存在某些区域样本对该算法很敏感而导致训练次数增加和“过拟合”现象出现。鉴于此，有学者提出充分利用样本的区域特性来构建区域分类器。如征荆等[5]提出将样本在特征空间聚类成不同区域，选择距测试样本最近区域的最优分类器组作为最后判别的分类器组，并取得较好效果。尽管这些研究都充分利用了样本区域特性建立了区域分类器，但还没充分考虑不同分类算法的互补性，以及各分类器集成的最优权重。

目前，基于成本敏感学习理论的电信客户流失预测研究[1~3]和基于模型集成理论的电信客户流失预测研究[4]都已出现，但还未见到同时含有这 2 类理论的研究文献。本文根据不同分类算法对不同区域样本具有不同适应性和不同分类算法之间存在一定互补性，提出了一种新的基于多分类器动态选择与成本敏感优化集成的电信客户流失预测模型。实验结果表明，所提出的客户流失预测方法和模型是可行且有效的。

1 理论基础

1.1 电信客户流失预测理论

KEAVENEY[6]定义的电信行业客户流失是指客户不再重复购买或终止原先使用的服务。电信客户流失之所以能够预测主要基于以下假设：电信客户的消费行为和习惯在一定程度上影射在其历史消费记录中，且在一定时期内保持相对稳定[7]。电信客户流失预测是个二分类问题，可描述为：

$$C=\begin{cases} c, & f(X)\geqslant\lambda; \\ n, & f(X)<\lambda, \end{cases} \tag{1}$$

式中，C 为客户流失状态函数，分为 c（流失）和 n（不流失）；$f(X)$ 为客户流失预测模型，输出客户流失概率；X 为客户特征属性集；λ 为客户流失判断阈值。

1.2 分类算法

根据文献和数据分析，本文将采用 NaiveBayes 算法、多层感知机算法和 J48 算法来构建区域子分类器。NaiveBayes 算法是一种简单、有效的分类算法，具体算法见文献[8]。多层感知机算法是神经网络中的一种分类算法，它在许多领域都得到应用，具体算法见文献[9]。J48 算法是属于决策树分类算法之一，它的分类效果较突出。

1.3 分类器成本敏感学习理论

目前分类器成本敏感学习理论是数据挖掘和机器学习中的前沿课题，我国在该领域刚刚起步，因而文献较少。该理论最早可以追溯到 1984 年 BREIMAN 等[10]对分类回归树的研究；随后 DOMINGOS[11]提出了基于 Bagging 的 MetaCost 算法；TING[12]提出了代价敏感的决策树算法；GAMA[13]提出了基于朴素贝叶斯的代价敏感学习。目前分类器成本敏感学习理论多用于模式识别和故障诊断领域，而在电信客户流失预测领域则非常稀少[1~3]。分类器成本敏感学习理论的实现主要有 3 种模式：改变样本分布、修改算法结构和修改预测结果。目前最常见模式是前 2 种，第 3 种模式极为少见。由于本文将采用多分类器动态集成模式，因而只能选择第 3 种实现模式来构建预测模型。

1.4 多分类器动态集成理论

从 HANSEN 等[14]于 1990 年开创性地提出了神经网络集成后，很多研究人员对多分类

器集成理论基础进行了探讨，将其应用到实际问题域中，并取得了很好的效果[15,16]。多分类器集成方法有线性集成、非线性集成和动态集成。前2种比较常见，动态集成研究却不多。多分类器动态集成又可分为：多分类器动态选择、多分类器动态集成和多分类器动态选择与动态集成。

2　基于多分类器动态选择与成本敏感优化集成的电信客户流失预测模型

针对电信客户流失预测具有成本敏感性和非对称性特点，因此需要采用基于成本敏感学习理论和多分类器动态集成理论来构建电信客户流失预测模型。在构建模型的过程中，重点解决3个问题：①如何实现多分类器的动态集成？②如何建立多分类器的成本敏感学习？③如何求得多分类器的最优组合权重？

2.1　集成模型的基本原理

本文将采取多分类器动态选择与优化集成的思想来建立客户流失预测集成模型（见图1）。

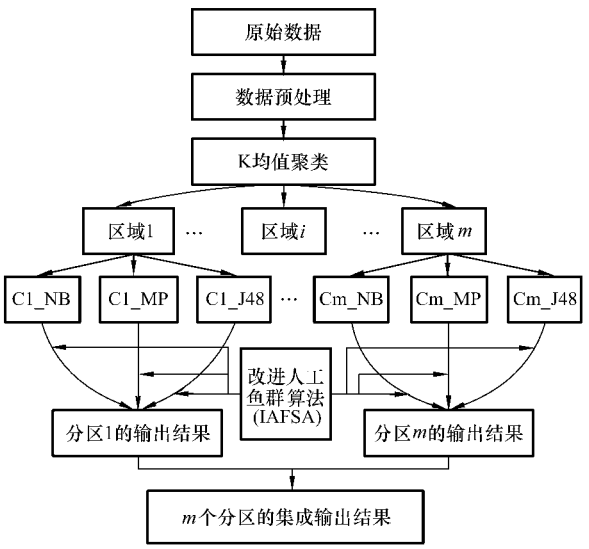

图1　基于多分类器动态选择和成本敏感优化集成的客户流失预测模型

（）集成模型训练过程的基本原理　①将原始数据进行预处理后，使用K均值聚类算法将训练集的样本在特征空间中聚类成个不同的分区，在相同分区内的样本具有最大的相似性，而相异分区间的样本具有最大的相异性；②基于每个分区中的样本分别使用差异很大的NaiveBayes算法（NB）、多层感知机算法（MP）和J48算法（J48）构建隶属于各个分区的客户流失预测子分类器；③构建以每个分区中多分类器线性组合预测利润函数为优化目标的优化决策问题，并使用改进人工鱼群算法（IAFSA）分别求解每个分区中的优化决策问题，以此求得多分类器的线性组合最优权重系数。

（）该集成模型测试过程的基本原理　①经过数据预处理的测试样本通过由K均值聚类法训练的分类器判断出样本的归属区域；②使用该区域的NB子分类器、MP子分类器和J48子分类器对测试样本进行分别预测；③使用IAFSA优化得到的线性组合最优权重系数将各子分类器的预测结果进行集成得到测试样本的最后预测结果。

根据集成模型训练过程和测试过程分析得到：①通过由K均值聚类算法训练的分类器对样本归属进行分类识别，这样就解决了样本动态选择各分区归属问题，实现了多分类器的动态选择。这种模式充分利用了各分区的样本区域特性，进而使各分区分类器具有更好的分类性能。②使用基于错误预测成本的成本敏感学习利润函数作为集成模型构建函数和评价标准，从而使构建的客户流失预测集成模型具有更好的预测效果和更强的适应性。③当各分区多分类器的线性组合权重经由IAFSA优化后，则实现了多分类器的优化集成，这样就充分利用了不同分类算法的互补性使集成模型的预测性得到进一步提高。

2.2　多分类器动态选择

在集成模型中，实现多分类器动态选择的关键是使用了K均值聚类算法[17]，它将客户样本在特征空间中聚类成不同的区域，并在不同区域中建立了多个不同分类器。因此若某客户样本落入不同区域，就可以使用对应区域分类器进行客户流失预测，从而实现多分类器动态选择。基于K均值聚类算法[17]的多分类器动态选择算法如下：

步骤1　输入预处理后的客户数据。

步骤2　使用K均值聚类建立数据分区域模型，

步骤2.1　选择$K(=m)$值以确定簇总数；

步骤2.2　在训练数据集中任意选择$K(=m)$个样本实例，把它们作为初始的簇中心；

步骤2.3　使用简单的欧氏距离将剩余样本实例赋给距离它们最近的簇中心；

步骤2.4　使用每个簇中的样本实例来计算每个簇的新平均值；

步骤2.5　如果新平均值等于上次迭代过程中的平均值，终止该过程，否则，使用新平均值作为簇中心，重复步骤2.3～步骤2.5。

步骤3　测试样本通过基于K均值聚类的

分区分类模型后，就分别进入各自的所属区域，并选择对应区域的子分类器进行预测。

2.3 多分类器成本敏感优化集成

当电信客户样本实现了多分类器动态选择之后，就可以将各区域分类器的预测结果进行成本敏感优化集成。电信客户实际状态的数学描述如下：

$$S_{\text{act},i}=\begin{cases}c, & p_{\text{act},j}=1;\\ u, & p_{\text{act},i}=0,\end{cases} \tag{2}$$

式中，$S_{\text{act},i}$ 为第 i 个客户的实际流失状态；$p_{\text{act},i}$ 为第 i 个客户的实际流失概率。当 $S_{\text{act},i}=c$ 时，则该客户的实际流失概率为 1；否则为 0。

当使用多分类器集成模式对电信客户进行预测其未来的流失状态时，其预测的数学描述如下：

$$S_{\text{pre},i}=\begin{cases}c, & \hat{y}_i\geqslant\alpha;\\ u, & \hat{y}_i<\alpha,\end{cases} \tag{3}$$

式中，$S_{\text{pre},i}$ 为第 i 个客户的预测状态；$\hat{y}_i$ 为集成模型对第 i 个客户的组合预测值；α 为预测判断阈值。若集成模型预测出客户流失概率大于或等于阈值，则该客户将流失；否则该客户不流失。

多分类器线性集成模型为

$$\hat{y}_i=\sum_{k=1}^{K}w_k y_{ik}, \tag{4}$$

式中，y_{ik} 为第 k 个子分类器对第 i 个客户的预测值，$k=1,2,\cdots,K$；w_k 为集成模型中第 k 个子分类器的组合权重系数，共 K 个子分类器。

由此，根据客户实际状况和客户流失集成预测结果将产生分类错差矩阵（见图 2）。

	预测流失	预测不流失
实际流失	N_{11} C_{11}	N_{10} C_{10}
实际不流失	N_{01} C_{01}	N_{00} C_{00}

参数说明：
① N_{10}表示把实际流失客户预测为不流失客户的数量，其他类似。
② C_{10}表示把一个实际流失客户预测为不流失客户的收益，其他类似。

图 2 客户流失预测模型的错差矩阵

在电信客户流失管理实践中，在图 2 中的 4 种预测收益是不相等的，通常有如下关系：

（1）正确预测收益均为正：$C_{11}>0$，$C_{00}>0$，$C_{11}\gg C_{00}$，即把一个实际流失客户正确预测为流失所获得的收益远远大于把一个实际不流失客户正确预测为不流失所获得的收益；

（2）错误预测收益均为负：$C_{10}<0$，$C_{01}<0$，$|C_{10}|\gg|C_{01}|$，即把一个实际流失客户错误预测为不流失客户造成的损失远远大于把一个实际不流失客户错误预测为流失客户而浪费的营销费用。

根据上面的分析，可以建立基于多分类器集成的电信客户流失预测利润函数

$$\max J=N_{11}C_{11}+N_{00}C_{00}+N_{10}C_{10}+N_{01}C_{01}, \tag{5}$$

式中，$(N_{11}C_{11}+N_{00}C_{00})$是正确预测而得到挽留机会的潜在收益；$(N_{10}C_{10}+N_{01}C_{01})$是错误预测而失去挽留机会的潜在损失；$C_{11}$、$C_{10}$、$C_{00}$、$C_{01}$是通过行业领域专家调查获得；$N_{11}$、$N_{10}$、$N_{00}$、$N_{01}$是通过训练模型的分类错差矩阵（见图 2）获得。

本文的多分类器集成是按照式(5)进行成本敏感学习的。此模型不仅考虑模型的正确预测收益，还考虑了错误预测损失，从而该模型在电信客户流失管理应用实践中较基于预测精度的方法更具有科学性和现实意义。

2.4 基于人工鱼群算法的集成模型权重系数求解

在线性集成中，如何寻找一组优化的组合权重系数是模型集成取得成功的关键。国内外学者在解决组合权重系数最优问题时，将更多的目光放在一些新的智能优化算法上面，而人工鱼群算法（AFSA）就是智能优化算法中较为年轻的一种，它是我国学者李晓磊等[18]于 2002 年提出的一种群集智能随机优化算法，是模拟鱼群的觅食、聚群和追尾等行为，从构造单条鱼的行为做起，通过鱼群中各个体的局部寻优达到全局最优的目的。该算法具有全局搜索、快速收敛等优点，因此，本文将首次引进人工鱼群算法来解决电信客户流失预测集成模型的组合权重系数的优化决策问题。

AFSA 的相关定义和行为描述[18,19]：

（1）相关定义　人工鱼个体的状态可表示为向量 $\boldsymbol{X}=[x_1,x_2,\cdots,x_n]$，$x_i$ 为寻优变量；人工鱼在当前位置的食物浓度为 $F=f(\boldsymbol{X})$；人工鱼个体之间的距离 $d_{ij}=\|\boldsymbol{X}_i-\boldsymbol{X}_j\|$；$V$ 为人工鱼的感知距离；e 为人工鱼的移动步长；δ 为人工鱼拥挤度因子。

（2）人工鱼的行为描述　觅食行为、聚群行为和追尾行为。

(i) 觅食行为。当人工鱼的状态为 $\boldsymbol{X}_i$，在可见域 $d_{ij}\leqslant V$，随机选择一个状态 $\boldsymbol{X}_j$，如果 $\boldsymbol{X}_i<\boldsymbol{X}_j$，则向该方向前进一步；反之，则随机重新选择状态，判断是否满足条件；如此反复几次后仍不满足条件则随机前进一步。其数学描述如下：

$$\begin{cases}x_{\text{inext}\,k}=x_{ik}+\dfrac{R(e)(x_{jk}-x_{ik})}{\|\boldsymbol{X}_j-\boldsymbol{X}_i\|}, & F_j>F_i;\\ x_{\text{inext}\,k}=x_{ik}+R(e), & F_j<F_i,\end{cases} \tag{6}$$

式中，$i=1,2,\cdots,n$；$R(e)$为$[0,e]$间的随机数；

x_{ik}、x_{jk}和$x_{\text{inext}\,k}$分别为状态向量$\boldsymbol{X}_i$、$\boldsymbol{X}_j$和$\boldsymbol{X}_{\text{inext}\,k}$的第$k$个元素。

(ii)聚群行为。人工鱼的当前状态为$\boldsymbol{X}_i$，在可见域内的伙伴数目为n，形成集合K_i，且

$$K_i=\{\boldsymbol{X}_j \mid \|\boldsymbol{X}_j-\boldsymbol{X}_i\|\leqslant V\}。\tag{7}$$

若$K_i\neq\phi$，则表示可见域内存在其他伙伴，则按下式探索伙伴的中心位置X_c。

$$X_{ck}=\frac{1}{n}\sum_{j=1}^{n}X_{jk},\tag{8}$$

式中，X_{ck}表示中心位置向量$\boldsymbol{X}_c$的第k个元素；X_{jk}表示第j个伙伴$\boldsymbol{X}_j$的第k个元素。计算该中心位置食物浓度F_i，如果满足

$$\mathrm{e}^{A_q}\left(\frac{F_i}{\delta}\right)>F_i,\quad A_q\leqslant 1,\quad \delta>1,\tag{9}$$

表明伙伴中心食物浓度较高，且不拥挤，则执行下式；否则执行觅食行为。

$$x_{\text{inext}\,k}=x_{ik}+\frac{R(e)(x_{ck}-x_{ik})}{\|\boldsymbol{X}_c-\boldsymbol{X}_i\|}。\tag{10}$$

若$K_i=\phi$，则表示可见域内不存在其他伙伴，则执行觅食行为。

(iii)追尾行为。人工鱼的当前状态为$\boldsymbol{X}_i$，在可见域内所有伙伴中最大的伙伴$\boldsymbol{X}_{\max}$，满足

$$F_{\max}=\delta F_i。\tag{11}$$

表明伙伴$\boldsymbol{X}_{\max}$的食物浓度高且不拥挤，则执行下式；否则执行觅食行为。

$$x_{\text{inext}\,k}=x_{ik}+\frac{R(e)(x_{\max k}-x_{ik})}{\|X_{\max k}-X_i\|},\tag{12}$$

式中，$X_{\max k}$表示状态向量$\boldsymbol{X}_{\max}$的第k个元素。

若人工鱼在当前可见域内无其他伙伴，则执行觅食行为。

(iv)公告板　算法中设置一个公告板用以记录人工鱼个体的最优状态和该位置的食物浓度。每条人工鱼执行一次操作后将自身状态与公告板进行比较，若优于公告板，则用自身状态取代公告板状态。

基本人工鱼群算法在解决实际问题时还有一些不足：如人工鱼步长e与视野V对算法的收敛速度和收敛精度影响很大。若设置不当则会陷入局部极值或者达不到精度[19]。针对以上不足，本文利用文献提出的改进人工鱼群算法如下：

(i)变尺度步长[9]　人工鱼可以根据当前的环境恶劣程度调整移动的步长。变步长

$$e_{\text{inext}}=\frac{(F_{\max}-F_i)}{(F_{\max}-F_{\min})}e,\tag{13}$$

式中，F_i为当前食物浓度；$F_{\max}$为在视野内的最大食物浓度；$F_{\min}$为在视野内最小食物浓度。

步长e的每次迭代都从环境中充分获得了有用信息——食物浓度，并利用其对自身进行改进，在迭代初期迭代速度很快，但随迭代的进行，步长会逐渐减小，有利于前期加快搜索进度，且后期提高局部搜索精度。

(ii)自适应视野[9]　人工鱼群的视野也随迭代过程进行自适应改变，这有利于人工鱼群的寻优。自适应计算公式如下：

$$V_{\text{inext}}=V_{\max}-\frac{(V_{\max}-V_{\min})}{i_{\max}}k,\tag{14}$$

式中，$V_{\max}$、$V_{\min}$分别为视野的最大值和最小值；$i_{\max}$为最大迭代次数；k为当前迭代次数。

在寻优初期，每条人工鱼在较大的视野内游动，扩展了算法的搜索范围，后来逐渐减小，使鱼能在缩小的视野内进行更细致的寻优。

(iii)改进觅食行为[9]　随机移动若干次，如果有改善则向更好的方向游去，按照概率P向全局最优值移动一步；否则按照概率$1-P$随机选择下一个状态，其计算式为：

$$\begin{cases}x_{\text{inext}\,k}=x_{ik}+\dfrac{R(e)(x_{jk}-x_{ik})}{\|\boldsymbol{X}_j-\boldsymbol{X}_i\|}, & \text{按概率 } P;\\ x_{\text{inext}\,k}=x_{ik}+R(e), & \text{按概率 } 1-P。\end{cases}\tag{15}$$

这3个改进策略都是对寻优过程进行优化，其中前面2个策略都是在接近最优解时调节寻优的速度和步伐，避免在函数奇异值的地方寻优失败，最后一个改进策略是为了避免陷入局部最优区域而错失全局最优解。经过这3种改进策略的处理能加快寻优速度，提高全局寻优精度。

改进人工鱼群算法在解决式(5)型问题时，需要将约束优化问题转化为非约束优化问题。此外，由于改进人工鱼群算法是求最大值问题，因此，基于多分类器动态选择多区域后的成本敏感集成模型的组合权重系数优化问题可更新为：

$$\max J'_i=J_i-A\Big(1-\sum_{k=1}^{3}w_{ki}\Big)^2,\tag{16}$$

$$\text{s.t.}\quad 1\geqslant w_{ki}\geqslant 0,\ k=1,2,3;\ i=1,2,\cdots,m,$$

式中，J'_i为区域i的优化目标函数；J_i为区域i的预测模型的预测利润；w_{ki}为区域i内预测模型k的组合权重系数；i为区域序号；A为等式约束条件的惩罚因子，一般是一个较大常数。当等式约束条件满足时，式中的最后部分就为0；否则就是一个较大的数，这样就能够实现式(5)型优化问题的解决。

使用前面介绍的改进人工鱼群算法求解式(16)的优化问题时，得到集成模型的最优组合权重系数w_{ki}^*后，再将w_{ki}^*代入式(4)即可得到客户流失预测集成模型对第i个客户流失概率

的组合预测值，最后将 $\hat{y}_i^*$ 代入式(3)求出集成模型对第 i 个客户流失状态的判断。

2.5 集成模型的评价

本文模型评价标准的计算式为

$$O = N_{11}C_{11} + N_{00}C_{00} + N_{10}C_{10} + N_{01}C_{01}, \quad (17)$$

式中各参数的说明见式(5)。

3 实证分析

3.1 实验数据

限于和合作公司的保密协议约束，本文涉及的实验数据不宜作更详细说明。实验所用原始数据为某电信企业某年 1～7 月的20 000个语音客户数据，这些客户不仅在网时长都超过 1 年时间，且 6 月份都在网，但有部分客户 7 月份处于离网状态。经过行业领域专家参与数据属性选择，最后得到的客户数据包括 132 个原始数据属性，主要由客户注册登记数据、客户通话行为数据、客户缴费行为数据、客户费用结构数据、客户服务投诉数据等组成。首先对原始数据进行清洗处理；接着对样本中的连续属性值使用基于 SOM 神经网络进行非监督式的离散化处理；然后使用粗糙集属性约简法对离散属性进行约简，最后获得一个含 8 个属性的最小约简属性集用于实证分析。在实证中，将 1～6月的客户数据作为输入指标，7 月的客户状态作为输出指标进行建模，而测试数据也是从该数据集中提取的，但不重复。各数据集结构见表 1。

表 1 各数据集结构

数据集名称	客户类型	数量	比例/%
原始数据集	流失客户	626	3.13
	非流失客户	19 374	96.87
训练集 D_{Train}	流失客户	225	3.22
	非流失客户	6772	96.78
测试集 D_{Test}	流失客户	89	2.94
	非流失客户	2937	97.06

3.2 实验结果

为了验证本文所提出的基于多分类器动态选择与成本敏感优化集成的电信客户流失预测模型的有效性，根据图 1 的模型原理设计如下实验：

实验 1 基于 K 均值的特征空间聚类实验 该实验主要是依据样本特征空间的聚类分布区域的不同，首先采用 K 均值聚类算法在训练集 D_{Train} 上进行聚类分析，获得一个基于样本特征空间 K 均值聚类的、用于样本区域识别判断的分类器，以及 K 个互不相同的训练集(D^1_{Train}、D^2_{Train}、D^3_{Train})，再使用该分类器对测试集 D_{Test} 进行分类测试，从而将测试样本分成 K 个互不相同的测试数据集(D^1_{Test}、D^2_{Test}、D^3_{Test})，这就为后续动态选择不同区域的多分类器实验做了前提准备，该实验结果见表 2。

表 2 基于 K 均值的特征空间聚类实验

数据集	区域 1		区域 2		区域 3	
	训练集 D^1_{Train}	测试集 D^1_{Test}	训练集 D^2_{Train}	测试集 D^2_{Test}	训练集 D^3_{Train}	测试集 D^3_{Test}
样本数量	2 987	1 318	1 042	440	2 968	1 268
非流失客户数	2 970	1 317	1 003	432	2 799	1 188
流失客户数	17	1	39	8	169	80

注：①该实验是在数据挖掘软件 Clementine 12.0 的环境下实现的；②表中数据集关系为：$D_{\text{Train}} = D^1_{\text{Train}} + D^2_{\text{Train}} + D^3_{\text{Train}}$；$D_{\text{Test}} = D^1_{\text{Test}} + D^2_{\text{Test}} + D^3_{\text{Test}}$。

实验 2 基于人工鱼群优化集成模型对比实验 该实验分别使用 NaiveBayes 分类算法、多层感知机分类算法和 J48 分类算法在训练集 D_{Train} 上训练 3 个子分类器(NB 子分类器、MP 子分类器、J48 子分类器)，并使用改进人工鱼群算法对 3 个子分类器的输出结果进行优化集成，该实验结果见表 3。

表 3 基于人工鱼群优化集成模型对比实验

模型类型	数据集	模型名称	训练集预测利润	测试集预测利润
子分类器模型	D_{Train}	NB 模型	151 190	63 890
	D_{Train}	MP 模型	**152 440**	**66 490**
	D_{Train}	J48 模型	143 190	63 940
集成模型	3 个子模型的集成结果	基于改进人工鱼群算法优化集成模型	**158 490**	**67 240**

注：基于人工鱼群优化集成模型的组合权重系数为：NB 模型为0.534 5，MP 模型为0.348 3，J48 模型为 0.172 1。

实验 3 基于分类器动态选择与人工鱼群算法优化集成模型对比实验该实验是在实验 1 的基础上，在每个不同分区的训练数据集(D^1_{Train}、D^2_{Train}、D^3_{Train})上分别使用 NB 分类算法、MP 分类算法和 J48 分类算法构建基于每个分区的子分类器(D^1 NB 子分类器、D^1 MP 子分类器、D^1 J48 子分类器、D^2 NB 子分类器、D^2 MP 子分类器、D^2 J48 子分类器、D^3 NB 子分类器、D^3 MP 子分类器、D^3 J48 子分类器)，然后采用改进人工鱼群算法分别优化集成每个分区的子分类器的预测结果，并得到每个分区的集成模型和子分类器集成的最优组合权重，接着使用每个分区的测试集去测试基于各分区的集成模型，这样通过动态选择不同分区的分类器进行预测，并结合改进人工鱼群算法优化集成分区子分类器的预测结果，最后整理得到测试集 D_{Test} 的预测结果，该实验结果见表 4。

表 4　基于分类器动态选择与人工鱼群算法优化集成模型对比实验

模型类型	数据集	集成模型的组合权重系数[NB模型,MP模型,J48模型]	训练集预测利润	测试集预测利润
分区域优化集成模型	区域1	[0.007 4,0.694 6,0.298 0]	62 200	29 440
	区域2	[0.515 6,0.484 4,0.000 0]	22 660	8 290
	区域3	[0.694 0,0.275 1,0.030 9]	7618 0	3221 0
动态集成模型	3个区域综合结果	基于K均值和改进人工鱼群算法动态集成模型	161 040	69 940

注:3个区域综合预测结果中的训练集预测利润和测试集预测利润都是由3个区域的预测利润求和得到。

在3个实验中,实验参数设置如下:

(1)在K均值聚类算法中,$m = 3$;客户流失预测模型的判断阈值$\alpha = 0.5$;客户流失预测模型错差矩阵中的4个参数为:$C_{11} = 200$,$C_{10} = -100$,$C_{01} = -30$,$C_{00} = 20$;等式约束条件的惩罚因子$A = 10^{10}$。

值得说明的是在K均值聚类算法中,最先设定$m = 4$,但实验后却发现其中一个分区的流失客户数极为稀少,以至于不能进行后续的实验,因此就放弃$m = 4$。接下来设置$m = 2$,但实验发现由于聚类类别很少,也不能很好地区别样本,因此就放弃$m = 2$。最后设置$m = 3$,实验结果良好,因此最后取$m = 3$。一般来说,m的大小一定是与样本数量有关的。m的不同取值在一定程度上将会影响到预测结果,若m值越大,则样本的区域特征表现得就越充分,但是太大后则会出现区域分类器的泛化能力降低。本文是采用实验方法确定取值的。

(2)在改进人工鱼群算法中,变量维数为3,人工鱼的感知距离(视野)$V = 1.0$,最大感知距离(视野)$V_{\max} = 1.0$,最小感知距离(视野)$V_{\max} = 0.3$,人工鱼的初始移动步长$e = 0.2$,鱼群规模为100,可见域内的人工鱼群的伙伴数目为30,食物浓度连续无改进的最大迭代次数为50,改进人工鱼群算法的最大迭代次数为500。

3.3　实验分析

3.3.1　基于K均值的特征空间聚类实验

从表2的实验结果可以作如下分析:

()样本在3个特征聚类区域的分布有差异　从训练数据在3个特征聚类区域的分布来看,在区域1和区域3的分布几乎相当,但在区域2则少很多;从测试数据来看,在区域1和区域3的分布也大致相当,在区域2也少很多。因此,区域1和区域3是样本比较密集的区域,而区域2则较为稀疏。

ℓ)流失客户比率在3个特征聚类区域的分布有差异　在聚类区域1的训练集和测试集的流失客户比率分别为0.57%和0.08%,两者比值为7.13;在聚类区域2的训练集和测试集的流失客户比率分别为3.74%和1.82%,两者比值为2.05;在聚类区域3的训练集和测试集的流失客户比率分别为5.69%和6.31%,两者比值为0.90。由此可得如下结论:①训练集和测试集的流失客户比率由低到高的排列顺序是:区域1、区域2、区域3;②各区域训练集和测试集中客户流失比率的比值由大到小的排列顺序是:区域1、区域2、区域3。

综合上面的分析可以得到如下结论:①聚类区域1的样本分布密集,尽管在其区域上训练集和测试集的客户流失比率较小,但是在其区域上训练集和测试集的客户流失比率的比值却较大。②聚类区域2的样本分布稀疏,在其区域上训练集和测试集的客户流失比率居中,但是在其区域上训练集和测试集的客户流失比率的比值也居中。③聚类区域3的样本分布密集,尽管在其区域上训练集和测试集的客户流失比率较高,但是在其区域上训练集和测试集的客户流失比率的比值却较小。

3.3.2　基于人工鱼群优化集成模型对比实验

()3个子分类器模型之间的对比分析　在表3的3个子分类器模型中,使用MP分类算法训练所得到的子分类器模型在训练集上获得的预测利润(152 440)和在试集上所获得的预测利润(66 490)都是最大的。因此,MP模型是最优秀的模型。然而,尽管使用J48所获得的模型在训练集中获得的预测利润(143 190)低于NB模型所获得预测利润(151 190),但是它在测试集上所获得的预测利润(63 940)却高于NB模型所获得预测利润(63 890)。因此,实验结果表明,使用NB分类算法构建分类模型更容易产生过拟合现象,因而其泛化能力较低。

综合实验结果和上面分析可以得到:即使基于相同数据,由不同分类算法构建的分类器模型的预测能力也是有差异的。

ℓ)子分类器模型和集成模型的对比分析　从表3可得出,由改进人工鱼群算法集成由不同分类算法在相同数据集上构建的分类器模型而得到的集成模型比任何一个子分类器模型的预测效果都好。如集成模型比最好的子分类器模型(MP模型)在训练集和测试集上的预测利润分别高出3.82%和1.12%。实验结果表明,基于改进人工鱼群算法能有效提高集成模型的预测能力。

3.3.3 基于人工鱼群动态优化集成模型对比实验

(1)3 个区域预测利润对比分析 将表 4 中 3 个分区中的预测利润与对应区域的样本数相除得到的比率(预测利润/样本数)分别为:训练集(区域 1 为 20.823 6,区域 2 为 21.746 6,区域 3 为 25.667 1)、测试集(区域 1 为22.336 9,区域 2 为 18.840 9,区域 3 为25.402 2)。因此,在 3 个区域中,测试样本的预测利润由大到小的顺序为:区域 3、区域 1、区域 2,这与实验 1 中得到的结论(区域 1 和区域 3 是样本比较密集的区域,而区域 2 则较为稀疏)有一定的内在关联关系,即稀疏区域的样本预测较为困难,密集区域的样本预测相对容易。

(2)动态集成模型与其它模型的对比分析 比较表 3 和表 4,可以得到:①在实验 3 中,基于 K 均值聚类的多分类器动态选择与改进人工鱼群算法的动态集成模型比实验 2 中基于改进人工鱼群算法优化 3 个子分类器的集成模型的分类性能要好,如实验 3 中动态集成模型在训练集上的预测利润和测试集上的预测利润比实验 2 中优化集成模型分别高出 1.58%和 3.86%。②在实验 3 中的动态集成模型比实验 2 中最好的单模型在训练集上的预测利润和测试集上的预测利润分别高出 5.34%和 4.93%。实验结果说明,基于相同数据所建立的基于 K 均值聚类的多分类器动态选择与改进人工鱼群算法的动态集成模型不仅比任何单模型都有较为明显的分类优势,而且比传统意义上的多分类器优化集成模型也有较好的分类优势。

在实证中,通过改变式(6)的 4 个参数取值来测试参数的敏感性。限于篇幅,略去该过程。实验结果表明,这 4 个参数对于集成模型具有重要意义,它不仅是构成集成模型优化函数的重要参数,而且直接影响到最后集成模型的实证结果。但有 2 个问题值得注意:①这 4 个参数的取值一般是请业内专家估计获得;②这 4 个参数在同一电信企业变化不大,而不同企业则一般是不同的。

此外,本研究还对集成模型与常见的神经网络分类模型做了对比实验,最后实验结果表明:①集成模型的预测结果比神经网络的单一模型有更好的稳定性和更高的预测准确性。这是因为集成模型充分利用了样本的区域特性,以及不同分类算法之间的分类互补性;而神经网络模型训练是寻找一个满足样本特征整体空间的模型,因此神经网络模型的泛化能力和预测稳定性要差一些。②集成模型的训练时间远远小于训练一个预测精度较高的神经网络单一模型,尤其是当训练样本的数量较大时就尤为明显。

综上所述,本文提出基于 K 均值聚类的多分类器动态选择与改进人工鱼群算法的成本敏感优化集成方法和模型对于解决电信客户流失预测问题是有效且可行的,这为我国电信客户流失预测研究和客户挽留管理实践提供了一种新的思路和方法。

4 结语

本文针对电信客户流失问题的复杂性,提出了采用多分类器动态选择与成本敏感优化集成方法来代替传统意义上的单模型方法、以及基于预测精度的预测方法来解决电信客户流失预测问题。

本文实验结果证明:①基于多分类器动态选择与成本敏感优化集成的电信客户流失预测模型比任何单模型和单纯基于改进人工鱼群算法优化的集成模型在分类性能上都有明显的优势,这也说明本文所提出的多分类器动态选择与成本敏感集成方法对于解决电信客户流失预测问题是有效且可行的;②基于改进人工鱼群优化算法对解决电信客户流失预测多分类器优化集成问题具有一定优势。

本文提出的思路、方法和模型虽然是基于电信客户流失预测的,但稍加改进后即可方便地应用于其他领域。今后在此基础上,有必要将其他分类算法和最新智能优化算法融入到电信客户流失预测的动态集成模型研究中。

参 考 文 献

[1] 钱苏丽,何建敏,王纯麟. 基于改进支持向量机的电信客户流失预测模型[J]. 管理科学,2007,20(1):54～58.

[2] 蒋国瑞,司学峰. 基于代价敏感 SVM 的电信客户流失预测研究[J]. 计算机应用研究,2009,26(2):521～523.

[3] XIE Y Y, LI X, NGAI E W T, et al. Customer Churn Prediction Using Improved Balanced Random Forests[J]. Expert Systems with Applications, 2009, 36(3):5 445～5 449.

[4] 王纯麟,何建敏. 基于 AdaBoost 的电信客户流失预测模型[J]. 价值工程,2007(2):106～109.

[5] 征荆,丁晓青,吴佑寿. 基于最小代价的多分类器动态集成[J]. 计算机学报,1999,22(2):182～187.

[6] KEAVENEY S M. Customer Switching Behavior in

Service Industries: An Exploratory Study[J]. Journal of Marketing, 1995, 59(2): 71～82.

[7] MOZER M C, WOLNIEWICZ R. Predicting Subscriber Dissatisfaction and Improving Retention in the Wireless Telecommunications Industry [J]. Neural Networks IEEE Transactions, 2000, 11(3): 690～696.

[8] 刘丽珍，宋瀚涛，陆玉昌. 基于 NaiveBayes 的 CLIFNB 文本分类学习方法[J]. 小型微型计算机系统，2005，26(9)：1 575～1 577.

[9] 关健，刘大昕. 一种基于多层感知机的无监督异常检测方法[J]. 哈尔滨工程大学学报，2004，25(4)：495～498.

[10] BREIMAN L, FRIEDMAN J H, OLSEN R A, et al. Classification and Regression Trees[M]. Belmont: Wad-sworth International Group, 1984.

[11] DOMINGOS P. MetaCost: A General Method for Making Classifiers Cost-Sensitive[C]//Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Diego, CA, 1999: 155～164.

[12] TING K M. An Instance Weighting Method to Induce Cost-Sensitive Trees[J]. IEEE Transactions on Knowledge and Data Engineering, 2002, 14(3): 659～665.

[13] GAMA J. A Cost-sensitive Iterative Bayes[C]//DIETTERICH T, MARGINEANTU D, PROVOST F, et al. Workshop on Cost Sensitive Learning(ICML 2000). California: Standford University Press, 2000.

[14] HANSEN L K, SALAMON P. Neural Network Ena embles[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1990, 12(10): 993～1 001.

[15] 孙灏，杜培军，赵卫常，等. 基于多分类器组合的高分辨率遥感影像目标识别[J]. 地理与地理信息科学，2009，25:(1):32～35.

[16] 张石清，赵知劲. 基于多分类器投票组合的语音情感识别[J]. 微电子学与计算机，2008，25(12)：17～20.

[17] ROIGER R J, GEATZ M W. 数据挖掘教程[M]. 翁敬农，译. 北京：清华大学出版社，2003.

[18] 李晓磊，邵之江，钱积新. 一种基于动物自治体的寻优模式：鱼群算法[J]. 系统工程理论与实践，2002，22(11)：32～38.

[19] 曹承志，张坤，郑海英，等. 基于人工鱼群算法的 BP 神经网络速度辨识器[J]. 系统仿真学报，2009，21(4)：1 047～1 050.

（编辑　刘继宁）

通讯作者：罗彬(1974～)，男，四川渠县人。电子科技大学经济与管理学院(成都市　610054)博士研究生。研究方向为商务智能研究。E-mail:Luobin10000@163.com

2012 中国工程管理论坛征文通知

中国工程院主办的“中国工程管理论坛”旨在推动我国工程管理理论建设研究与提高工程管理实践水平，探讨我国工程管理现状和发展的关键问题，已成功举办了五届。

随着我国社会经济协调发展战略的不断深化以及金融危机对全球经济的深刻影响，加快转变经济发展方式正面临着前所未有的机遇和挑战，为此，中国工程院拟于 2012 年 9 月在合肥举办第六届中国工程管理论坛，主题为“加快转变经济发展方式与工程管理 ”。论坛诚邀学术界、产业界及社会各界的专家、学者参加，围绕经济发展方式转变相关工程管理问题以及中国工程管理理论体系等开展跨学科、跨行业、跨地区的学术研讨，为实现我国经济协调发展做出新的贡献。

现将论坛有关事项通知如下：

论坛主要议题：① 经济发展方式与工程管理；② 中国工程管理理论体系；③ 战略性新兴产业发展与工程管理；④ 节能降耗工程管理；⑤ (各)行业工程管理；⑥ 工程管理专业学位建设

主办单位：中国工程院　安徽省人民政府

承办单位：中国工程院工程管理学部　安徽省科技厅　中南大学　合肥工业大学

论坛时间和地点：时间：2012 年 9 月 21 日报到；2012 年 9 月 22～23 日　地点：合肥市稻香楼宾馆

论坛联系人：

联系人 1：王青娥　电话：0731-2655065　传真：0731-2655065　E-mail：cemf2012@vip.163.com

联系人 2：刘业政　电话：0551-2904991　传真：0551-2904991　E-mail：cemf2012@126.com

联系人 3：于泽华　常军乾　电话：010-59300345　010-59300280　传真：010-59300243

Email：yzh@cae.cn，cjq@cae.cn