面向数据漂移的代价敏感客户细分

邹 鹏，于 渤，王宪全

（哈尔滨工业大学 经济管理学院，150001 哈尔滨，zoupeng@hit.edu.cn）

摘 要：为解决数据挖掘中存在的数据漂移和客户价值分布不平衡问题，采用了分阶段聚类和代价敏感支持向量机的新方法.新方法首先对全部客户聚类得到特征相似的客户群，然后用某个区域客户属于某客户群的后验概率对城市进行聚类，具有相似后验概率分布的城市群被认为是具有类似的客户结构，每个城市群的客户组成了新的客户样本，对每个样本分别进行代价敏感分类，并完成客户细分.对比实验表明，该方法提高整体预测准确率和高价值客户识别能力，降低模型错误分类代价.改进的方法能在保证分类准确率的同时，更有助于企业锁定高端客户，动态地调整区域市场战略.

关键词：代价敏感学习；支持向量机；客户细分；数据漂移

中图分类号：TP18 **文献标志码**：A **文章编号**：0367－6234(2011)01－0119－06

Cost-sensitive learning method with data drift in customer segmentation

ZOU Peng, YU Bo, WANG Xian-quan

(School of Management, Harbin Institute of Technology, 150001 Harbin, China, zoupeng@hit.edu.cn)

Abstract: To solve the problem of data drift and asymmetric misclassification costs in customer segmentation, a cost sensitive learning method integrated with two-step cluster is proposed. This method firstly applied k-means cluster by the posterior probability distribution of give region to group similar regions together, and then used cost-sensitive support vector machine to find customer segmentation for each region-group. The results show that the cluster based on similarity of customer segmentation structure can improve the total accuracy and the proposed cost-sensitive support vector machine is an effective method to distinguish high value customers compared to the original support vector machine.

Key words: cost-sensitive learning; support vector machine; customer segmentation; data drift

数据挖掘的应用要面对各种复杂的数据分布，数据分布的特殊性在某些时间、空间条件下表现得更为突出.本文在对中国零售金融客户数据进行数据挖掘过程中发现了2种结构性的数据分布差异，这2种差异在数据挖掘的其他商业应用领域也有普遍存在.第1种差异是不同区域的客户人口结构差异.在人口统计学领域，这种随着经济环境、人口结构和生活方式的变迁和差异使样本人群的范围和特质发生变化被称为人口数据漂移(数据漂移)[1].数据漂移直接导致构建客户细分过程中不同样本之间及样本与总体之间的差异程度亦无法确定，限制了客户细分模型的适用性和推广性.第2种差异是客户价值分布的差异.客户赢利性研究发现，多数企业的前20%最有价值的客户创造的利润平均达到总利润的90%，而其他的客户不仅不能带来收益，还有可能带来亏损(利润为负值)[2].

这2种客户在数量上差距悬殊，高价值客户在全部客户中占比例很低，而普适性的数据挖掘方法是基于错误率的，在一个非平衡分布数据集上建立的模型只会对数量上占优的低价值客户有较强的识别能力，预测结果偏误向于大样本的低价值客户.根据这样的预测结果实施的市场策略不仅是把企业的资源消耗在无利可图的客户关系上，更严重的会造成高价值客户的流失.中国目前正处在经济和社会高速发展阶段，存在着地区间

收稿日期：2010－10－30.
基金项目：国家自然科学基金资助项目(70802019).
作者简介：邹 鹏(1975—)，男，讲师；
于 渤(1960—)，男，教授，博士生导师.

发展不平衡、居民收入和消费水平的差距过大. 就世界范围内存在的这 2 种消费者数据结构性差异现象来说,中国现阶段的表现尤为突出. 这 2 个问题严重地影响了客户细分的有效性,尤其是对高价值客户识别的精确性,针对这 2 个问题,本文分别对客户所在城市进行聚类和对客户进行分类,一方面解决这种多样本间差异不确定性,实现多区域客户的准确细分,另一方面引入代价敏感学习方法通过代价函数来自动修正分类偏误.

1 文献综述

1.1 数据漂移问题

数据漂移可分为时间上的漂移和空间上的漂移. 时间上的漂移指其他条件相同的样本人群的特征随时间推移发生变化;空间上的漂移则指其他条件相同的不同地理地区的样本人群特征的差异性. 本文研究的不同地区的客户的差异现象属于空间上的数据漂移[3],是由于地区间经济发展水平、人口构成和生活习惯差异造成的[4]. 客户细分是指企业在明确的战略、业务模式和特定的市场中,根据客户的属性、行为、需求、偏好以及价值等因素对客户进行分类[5]. 在涉及到不同地区客户细分时:1)在每个地区的客户样本上建立一个专门的预测模型,在某一样本上建立的模型用于预测其他样本,分类效果往往不如建模样本[6]. Michael 等[7]在建立纽约地区直销客户响应区分模型时发现,利用某地理区域客户样本数据建立的模型对其他地区客户分类出现较大误差. Bijmolt 等[8]在对 15 个欧洲国家客户进行实证研究时也发现了这种地区间的客户细分结果的差异. J. P. Li 等[9]在对客户进行信用评估时对每个城市的客户都单独建模,但是这种方法不仅消耗时间,效率低,而且无助于企业形成全局性的市场认知和可移植的市场策略;2)把客户样本按照地理分布划分为若干区域,再对每个区域样本建模,如把中国划分为华北、东北、华东和华南地区等. 这是一种基于主观判断或经验的划分,不一定真正反映消费者的真实差异,Mo 等[10]的研究发现经济和文化环境比地理属性更能影响消费者的行为. 尤其是在当前中国经济高速发展的阶段,不同地区间的经济形态,水平差距也在快速变化着,后发优势和区域产业升级已经使原有的经济地理版图发生了很大变化,如环渤海经济圈的形成就改变了传统的华北、华东和东北地区的经济区域格局,西部开发战略的实施拉动了四川和重庆这 2 个城市的经济起飞,静态的地理概念已难以反映变化的经济和社会生活变迁. 必须探索区域之间数据漂移的内在特征,建立新的客户样本空间划分方法. 本文首先根据客户分类的相似性把城市进行聚类,然后对每个城市群基础上形成的新的客户样本再进行细分.

1.2 客户价值差异

营销领域中的 20/80 理论指出企业最有价值的前 20% 客户为企业创造了 80% 的利润,在银行业,甚至是前 10% 的客户带来了 90% 的利润[11]. 所以把高价值客户错误划分为低价值客户造成的代价损失远远大于低价值客户错分的代价. 将低价值客户错分为高价值客户仅仅是浪费一些资源,而将高价值客户错分为低价值客户有可能造成高端客户的流失,甚至有损企业形象. 现有的多数技术都是优化分类的整体准确率[12],并没有区分这 2 种错误的不同,导致不能很好适应客户细分的要求. 有研究指出,涉及到产品评价和客户满意度评估的数据往往是非常不平衡的,在本文样本中高价值客户数量远远少于低价值客户. 现有的方法的预测结果会更倾向于个体数量较多的类别,即低价值客户,这样就会增加错误划分高价值客户的机会. 在商业领域有很多这样的类别不平衡数据,如大量的信用卡用户中极少数的有欺诈行为的客户[13],众多企业中有破产可能的少数公司[14]. 因为传统的基于整体准确率的数据挖掘技术不能很好地解决客户细分中错分代价不同的问题,本文尝试在支持向量机 (Support Vector Machine, SVM)基础上引入代价敏感学习,用错分代价最小化来评估分类效果.

代价敏感学习的研究可以追溯到 Granger[15]对文本内容分类的研究,Elkan[16]发展了 Granger 的研究,建立错分代价矩阵,将待预测样本分到期望代价最小的类别中. Turney[17]和 Kwedlo 等[18]建立了基于分类准确率和错分代价的拟合函数的遗传算法. 在客户细分方面,夏国恩等[19]用代价敏感学习预测电信客户流失,取得了一些有益的进展,但只是根据经验和文献主观设定错误分类的代价,不能根据客户数据的真实情况设定代价函数或取值,不能精确控制和优化分类预测的效果. SVM 是根据统计学习理论提出的一种新的机器学习方法,在解决小样本、非线性及高维模式识别问题中表现出许多特有的优势,有较好的泛化能力,而且当训练数据中各类样本数量差别较大时,可以通过参数设置调整偏置[20]. 本文尝试建立以代价最小化为评价标准的客户分类方法.

2　基于城市群聚类的代价敏感客户细分

基于城市群聚类的代价敏感客户细分方法首先用客户人口特征属性对全部客户聚类,得到若干特征相似的客户群,然后计算某个区域(城市)的客户属于某个客户群的后验概率. 再用后验概率对城市进行聚类,具有相似后验概率分布的城市可以被认为具有类似的客户分布结构[21],得到新的城市组划分,把每个城市组的客户组成新的客户样本,对每个新客户样本进行代价敏感分类,完成客户细分,如图 1 所示.

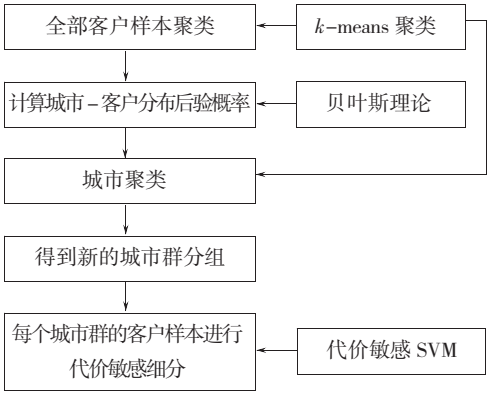

图 1　面向数据漂移的代价敏感客户细分方法过程

2.1　k-means 聚类

k-means 算法是一种基于样本间相似性度量的间接聚类方法. 此算法以 k 为参数,把 n 个对象分为 k 个簇,使簇内具有较高的相似度,而且簇间的相似度较低. 相似度的计算根据 1 个簇中对象的平均值来进行. 此算法首先随机选择 k 个对象,每个对象代表 1 个聚类中心. 对于其余的每 1 个对象,根据该对象与各聚类中心之间的距离,把它分配到与之最相似的聚类中,然后计算每个聚类的新中心. 重复上述过程,直到准则函数收敛,k-means 算法的工作过程和聚类中心个数确定见文献[22]. 在进行 k-means 聚类时采用文献[23-24]提出的优化准则. 即新增的类会提高聚类解释方差的比例,当解释方差提高的幅度开始显著下降,并且边际趋近于 0 时,停止增加新的聚类,此时得到的聚类为优化的结果.

2.2　计算城市-客户分布后验概率

设 P(cluster i, city j) 为第 j 个区域(城市)的客户属于第 i 个客户群的后验概率,它可以表示城市内部的客户细分结构的相似性[21],计算方法为:

1)计算 P(cluster i) = 第 i 个类的客户数量/全体客户总数;

2)计算 P(city j) = 第 j 个城市的客户数量/全体客户总数;

3)计算 P(city j, cluster i) = 第 j 个城市在第 i 个类中的客户数量/全体客户总数;

4)计算 P(city = j | cluster = i) = P(city j, cluster i)/P(cluster i);

5)根据贝叶斯理论计算后验概率 P(cluster i | city j).

2.3　基于城市-客户分布后验概率的城市聚类

本研究假设具有相似后验概率分布的城市会有相似的客户分类结构,所以利用该城市的客户属于某个客户群的后验概率以城市为个体进行聚类,得到若干城市群分组,每个城市群组里包括若干具有相似客户分类结构的城市. 将每个城市群组包含城市的客户合并为新的客户样本子集,这样得到若干以城市群为划分的客户样本集合.

2.4　代价敏感 SVM

本文在基本的 SVM 分类模型中引入代价敏感学习参数. 对二元分类问题,假定已知观测样本集(x_1, y_1), …, (x_i, y_i), …, (x_n, y_n), $x_i \in R^l$, $y_i \in \{+1, -1\}$, $i = 1, \cdots, n$ 能被超平面$(w \cdot x) - b = 0$ 分类,问题为最小化目标函数为

$$R(w,\xi) = \frac{1}{2}\|w\|^2 + C\Big(\sum_{i=1}^{n} \xi_i\Big),$$
$$\text{s.t. } y_i(x_i \cdot w + b) \geqslant 1 - \xi_i,\ \xi_i \geqslant 0, i = 1, \cdots, n. \tag{1}$$

目标函数中各参数意义和优化问题的求解见文献[25],得到最优分类函数为

$$f(x) = \text{sign}\Big\{\sum \alpha_i y_i k(x_i \cdot x) + b\Big\}. \tag{2}$$

式中 $k(x_i \cdot x)$ 为核函数,可以将低维输入空间中的线性不可分问题转化为高维属性空间中线性可分问题.

设样本集$(x_i, y_i, \cos t_i)$能被超平面$(w \cdot x) - b = 0$ 分类,那么问题为最小化目标函数为

$$R(w,\xi) = \frac{1}{2}\|w\|^2 + C\Big(\sum_{i=1}^{n} \cos t_i \xi_i\Big),$$
$$\text{s.t. } y_i(x_i \cdot w + b) \geqslant 1 - \xi_i,\ \xi_i \geqslant 0,\ i = 1, \cdots, n. \tag{3}$$

与式(1)不同,式(3)的经验代价考虑到不同样本的误差具有不同的误分类代价. 为求解优化问题式(2),使用上述方法得到改进的对偶 Lagrange形式为

$$L_D = \frac{1}{2}\sum_{i,j=1}^{n} \hat{\alpha}_i \hat{\alpha}_j y_i y_j (x_i x_j) - \sum_{i=1}^{n} \alpha_i,$$
$$\text{s.t. } \sum_{i=1}^{n} \alpha_i y_i = 0,\quad 0 \leqslant \alpha_i \leqslant \cos t_i C,\ i = 1, \cdots, n. \tag{4}$$

此时得到最优分类函数为

$$f'(x) = \operatorname{sign}\{\sum_{i,j=1}^{n} \hat{\alpha}_i y_i k(x_i \cdot x) + \hat{b}\}. \quad (5)$$

根据企业的财务数据和客户交易数据可以计算误分类代价函数或常数值,进一步应用改进的 SVM 在每个新的城市群客户样本建立预测模型.

3 实验与分析

3.1 数据

本文与国内某银行合作,通过电子邮件向 18 万客户发放问卷,最后搜集了全国 12 个城市的 13 382个客户的有效问卷,同时将这些客户的个人属性和交易记录进行集成. 将信息增益分析技术和基于多维数据分析的方法集成在一起删除信息量较少的属性,收集信息量较多的属性[26],最终选择了一些指标作为模型的相关属性集,包括年龄、收入、教育程度、信用额度、婚姻状况等.

3.2 数据实验

首先基于上述 5 个属性在全部客户样本上进行聚类,得到 5 个客户群,然后按照计算城市 - 客户分布后验概论方法计算每个城市的客户属于某个客户群的后验概率, 如表 1 所示(北京的客户属于客户群 1 的后验概率是 0.44). 利用表 1 的数据,以城市为对象聚类,得到城市分组,如表 2 所示.

表 1 城市 - 客户后验概率分布

城市	客户聚类				
	1	2	3	4	5
北京	0.44	0.16	0.18	0.05	0.17
上海	0.42	0.17	0.17	0.18	0.06
广州	0.45	0.17	0.15	0.18	0.05
深圳	0.22	0.40	0.18	0.07	0.13
南京	0.25	0.41	0.13	0.17	0.04
苏州	0.39	0.10	0.18	0.15	0.18
杭州	0.38	0.17	0.14	0.17	0.14
青岛	0.03	0.18	0.43	0.19	0.17
天津	0.06	0.19	0.42	0.18	0.15
武汉	0.23	0.40	0.16	0.07	0.14
重庆	0.19	0.41	0.07	0.17	0.16
成都	0.19	0.38	0.10	0.18	0.15

根据表 2 把客户按照新的城市分组合并成新的样本集合. 把客户按照为银行贡献利润分类,以平均利润为标准分为 2 类,同时根据每个客户的利润贡献设定动态的代价敏感函数,在每个样本上分别应用代价敏感 SVM 建立客户价值细分模型,样本集合为(x_i, y_i), 其中 $\boldsymbol{x}$ 为客户样本(多维向量), $y \in \{+1, -1\}$为高忠诚和低忠诚 2 类客户. 对于核函数和参数选择,本文用 MATLAB6.5 试验比较了多项式核函数 $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\delta \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{x}_j + r)^d$, sigmoid 核函数 $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \tan \mathrm{h}(\alpha \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{x}_j + \beta)$和径向核函数 $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\{-\lambda \mid \boldsymbol{x}_i - \boldsymbol{x}_j \mid^2\}$, 当 $\lambda = 0.01$时的径向核函数效果较好. 对 2 类别的分类,基于客户的交易数据和银行的财务数据计算式(3)中的错分代价矩阵,如表 3 所示,关系为 $\cos t_{A,B}$(A 错分为 B) $= 1/\cos t_{B,A}$(B 错分为 A), $\cos t_{B,A}$(B 错分为 A) $= 1/\cos t_{A,B}$(A 错分为 B).

表 2 城市分组样本量统计

城市分组	城市	城市样本量	城市组样本量
1	北京	1 336	6 082
	上海	1 333	
	广州	1 347	
	苏州	1 017	
	杭州	1 049	
2	南京	1 041	5 398
	武汉	1 057	
	重庆	1 029	
	成都	909	
	深圳	1 362	
3	天津	906	1 902
	青岛	996	

表 3 分类代价矩阵

客户类别	分为 A	分为 B
A 类客户	0	10.62
B 类客户	0.32	0

实验产生的分类结果中各种样本的数量如表 4所示.

表 4 分类结果

客户类别	分为 A	分为 B
A 类客户	a	b
B 类客户	c	d

表 4 中:

1)将 A 类客户分为 B 类客户的错误比率为

$$F_{A,B\cup C} = b/a + b.$$

2)将 B 类客户分为 A 类客户的错误比率为

$$F_{B,A} = c/c + d.$$

3)整体准确率为

$$T = (a + d)/(a + b + c + d).$$

4)期望损失函数定义为[27]

$$C = b * \cos t_{A,B} + c * \cos t_{B,A}.$$

5)平均错分代价为

$$AVE(C) = \frac{(b) F_{A,B} * \mathrm{cost}_{A,B} + (c) F_{B,A} * \mathrm{cost}_{B,A}}{a + b + c + d}.$$

最后用分类树算法提取各类客户的特征. 城市分组 1 的客户分类特征描述如表 5 所示. 这些规则可以用于该城市群组内新的客户数据,用来预测其价值.

表 5　城市分组 1 的客户价值分类特征规则

客户分类	特　　征
A 类(高价值)	√ 年龄 <25,未婚,年收入 80 000 ~ 120 000. √ 年龄 >25,已婚,年收入 >120 000, 教育程度为大学. √ 年龄 >25, 已婚,年收入 80 000 ~ 120 000.
B 类(低价值)	√ 年龄 <25,未婚,信用额度 <5 000. √ 年龄 < 25, 已婚,收入 <80 000. √ 年龄 > 25 未婚,收入 80 000 ~ 120 000, 教育程度为高中以下.

3.3　对比分析

为了检验提出方法的先进性,进行了实验对比分析,设计了 2 种对比方法:1)按经验划分城市群组的基本 SVM 方法;2)按后验概率聚类城市群组的基本 SVM 与后验概率聚类城市群组的代价敏感 SVM 比较. 按经验划分是指按照地理区域把 12 个城市划分为华北、华东、华南、华中和西南 5 个地区,方法 1 和 2 分别在 3 个城市群组上建立分类模型. 实验过程首先把每个城市组的客户数据都按照 2∶1 分为训练集和测试集,在训练集上建立分类模型,在测试集上测试准确率和错误分类代价等指标,对比结果如表 6 所示. 方法 1 在按照传统地理区域划分的客户数据集建模,分类准确率较低,这是因为在每个地理区域样本中,客户分布的结构实际上差异是较大的,在训练数据集上建立的模型对测试数据集拟合效果较差,另外虽然将 96% 的高价值客户错误分为低价值,但由于高价值客户数量很少,所以总体准确率仍然较高,这样不能识别绝大多数高价值客户,对客户价值细分和高端客户管理是没有意义的. 方法 2 的实验过程中,由于重新分组的城市群内部客户结构相似度较高,训练集上建立的模型也会更好的拟合测试集,高价值客户正确识别率没有提高,平均代价下降也不显著. 本实验不仅将 96% 的高价值客户正确识别,同时能将其他类别客户的正确识别率也控制在较好的范围,同时显著降低平均代价. 结果证明,在按照客户结构相似度划分城市群的客户数据集上基于代价敏感学习机的 SVM 比传统的 SVM 能在保证分类准确率的同时,显著降低模型分类的代价,显著提高了高价值客户的识别能力.

表 6　对比试验结果

方法	T^*/%	$F^*_{A,B\cup C}$/%	$F^*_{B\cup C,A}$/%	$AVE(C)^*$
1	65.7	96.0	3.0	1.46
2	93.2	95.4	5.7	1.22
3	93.0	4.0	5.6	0.06

注:* 为在各样本建模分类预测的平均值.

4　结　论

1)根据客户特征相似性将城市重新组合,在得到的新城市群组上建立细分模型. 数据实验和对比分析表明,提出的方法提高了客户细分的整体准确率,有助于动态地调整区域市场战略.

2)数据实验和对比分析表明,提出的方法不仅提高了精确识别高价值客户的能力,有助于企业锁定高端客户,实施精准营销.

参考文献:

[1] 郭尕,梁世栋,方兆本. 消费者信用评估分析综述[J]. 系统工程, 2001, 19(6): 9 - 15.

[2] SELDEN L, COLVIN G. Angel customers and demon customers: discover which is which and turbocharge your stock [M]. Portfolio Hardcover. New York: Springer, 2003: 20 - 30.

[3] ARBIA G, LAFRATTA G. Evaluating and updating the sample design in repeated environmental surveys: monitoring air quality in padua[J]. Journal of Agricultural Biological and Environmental Statistics, 1997, 8(4): 451 - 466.

[4] FERGUSON C C. Techniques for handling uncertainty and variability in risk assessment models [M]. Berlin: Umweltbundesamt, 1998: 160 - 186.

[5] LIAUTAUD B, HAMMOND M. 商务智能 [M]. 北京:电子工业出版社, 2002:186 - 205.

[6] LI Yijun, ZOU Peng, YE Qiang. Customer sample differenceoriented bayes segmentation algorithm [C]//Proceedings of 2006 International Conference on Management Science and Engineering. New York: IEEE, 2006:3 - 7.

[7] MICHAEL J A B, LINOFF G S. Data mining techniques: for marketing sales, and customer relationship management [M]. New York: JohnWiley & Sons, Inc, 2004: 120 - 150.

[8] BIJMOLT T H T, PAAS L J, VERMUNT J K. Country and consumer segmentation: mult-level latent class analysis of financial product ownership [J]. International Journal of Research in Marketing, 2004, 21(4): 323 - 340.

[9] LI J P, XU W X, SHI Y. Credit scoring and principal component analysis linear-weighted comprehensive as-

sessment and application [J]. System Engineering, 2004, 18(1): 64 - 68.

[10] MO J H, ZOU P, LI Y J. Context of the concept drift in data mining: an empirical study on the regional economic influence to the relation between demographic attributes and credit card holder loyalty [C]//Proceedings of 2008 International Conference on Management Science and Engineering. New York: IEEE, 2008: 38 - 43.

[11] BLOOMQUIST A. Small business, big profits [J]. Bank Marketing, 1998, 7(8): 18.

[12] ZHANG Y, REN Y Y, ZHANG Z Y, *et al*. Fuzzy SVDD classifier with positive and negative samples [J]. Journal of Computational Information Systems, 2008, 4(3): 1055 - 1062.

[13] SYEDA M, ZHANG Y Q, PAN Y. Parallel granular neural networks for fast credit card fraud detection [C]//Proceedings of the 2002 IEEE International Conference on Fuzzy Systems. New York: IEEE, 2002: 30 - 38.

[14] TAM K Y, KIANG M Y. Managerial applications of neural networks: the case of bank failure predictions [J]. Management Science, 1992, 38 (7): 926 - 947.

[15] GRANGER C W J. Prediction with a generalized cost of error function [J]. Operational Research Quarterly, 1969, 20(2): 199 - 207.

[16] ELKAN C. The Foundations of costsensitive learning [C]//Proceedings of the 17th International Joint Conference on Artificial Intelligence. San Francisco, CA: Morgan Kaufmann Publishers Inc., 2001: 973 - 978.

[17] TURNEY P D. Costsensitive classification: empirical evaluation of a hybrid genetic decision tree algorithm [J]. Journal of Artificial Intelligence Research 1995, 23(2): 369 - 409.

[18] KWEDLO W, KRETOWSKI M. An evolutionary algorithm for cost-sensitive decision rule learning [C]// Proceedings of the 12th European Conference on Machine Learning. London, UK: Springer-Verlag, 2001: 288 - 299.

[19] XIA Guoen, JIN Weidong. Tradeoff of errors of two types in customer churn prediction [J]. Journal of Marketing Science, 2006, 2(4): 1 - 7.

[20] VAPNIK V N. Statistical learning theory [M]. New York: Wiley, 1998: 54 - 62.

[21] MO Jiahui, KIANG M Y, ZOU Peng, *et al*. A two-stage clustering approach for multi-region segmentation [J]. Expert Systems with Applications, 2010, 37 (10): 7120 - 7131.

[22] DAVISON I. Understanding *k*-means non-hierarchical clustering [EB/OL]. [2002 - 10 - 03]. http://www.cs. albany. edu/ ~ davidson/courses/CSI635/Understanding K-MeansClustering. pdf.

[23] ALDENDERFER M S, BLASHFIELD R K. Cluster analysis [M]. Newbur Park, CA: Sage, 1984: 18 - 24.

[24] DAVID J, Jr KETCHEN, CHRISTOPHER L. The application of cluster analysis in strategic management research: an analysis and critique [J]. Strategic Management Journal, 1996, 17(6): 441 - 458.

[25] CLARKE D W, MOHTADI C, TUFFS P S. Generalized predictive control [J]. Automatic, 1987, 23(2): 137 - 148.

[26] HAN Jiawei, KAMBER M, PEI Jian. Data mining: concepts and techniques [M]. Second Edition. San Mateo: Morgan Kaufmann Publishers Inc, 2001: 172 - 176.

[27] JOOS P, VANHOOF K, OOGHE H. Credit classification: a comparison of logit models and Decision [C]// Proceedings of ECML Workshop on Applications of Machine Learning and Data Mining in Finance. Forschung, Germany: Technical University of Chemnitz, 1998: 59 - 73.

(编辑 张 红)