计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2017,26(11):170−175 [doi: 10.15888/j.cnki.csa.006078]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于 RFM 模型的半监督聚类算法①

程汝娇, 徐鸿雁

(西南财经大学天府学院, 绵阳 621000)

摘　要: 客户分类作为客户关系管理 (CRM) 的重要管理方法, 是企业进行市场营销的重要依据. 通过对客户进行分类, 有利于对客户价值进行准确评估, 方便进行精准营销. 本文通过对 RFM 模型数据集本身潜藏的先验结构化信息进行研究, 标记出两组客户数据作为先验类别标记, 进而得到两个初始聚类中心. 基于传统 K-means 算法使用自适应方法确定 K 值和初始聚类中心. 引入 Must-link 和 Cannot-link 两种约束将类别标记转换为成对约束信息, 基于 HMRF-KMeans 成对约束, 引入约束惩罚项和约束奖励项, 实现对聚类引导和聚类结果的调整. 使用改进的半监督聚类算法 (RFM-SS-means) 对标准数据集进行了测试, 同时使用 Food mart 数据集对比了 RFM-SS-means 算法与传统 K-means 算法、two-steps 算法的聚类效果. 由实验结果可知, RFM-SS-means 的 CH 系数最大, 无需事先确定 K 值和初始聚类中心, 聚类效果良好.

关键词: 客户分类; 半监督聚类; K-means; RFM 模型

引用格式: 程汝娇,徐鸿雁.基于 RFM 模型的半监督聚类算法.计算机系统应用,2017,26(11):170–175. http://www.c-s-a.org.cn/1003-3254/6078.html

Semi-Supervised Clustering Algorithm Based on RFM Model

CHENG Ru-Jiao, XU Hong-Yan

(Tianfu College of Southwest University of Finance and Economics, Mianyang 621000, China)

Abstract: As an important management method of customer relationship management (CRM), the customer classification is the basis for enterprises to carry out marketing. The classification of customers is conducive to accurate assessment of customer value and facilitate the precise marketing. In this paper, we study the priori structured information hidden in the RFM model dataset, and mark two sets of customer data as a priori category mark, and then get two initial clustering centers. Based on the traditional K-means algorithm, the K value and the initial clustering center are determined with the adaptive method. Combining the two types of constraints of Must-link and Cannot-link, the category markers are transformed into pairs of constraint information. Based on HMRF-KMeans pairs, the constraints and constraint bonuses are introduced to improve the clustering guidance and clustering results. The improved semi-supervised clustering algorithm (RFM-SS-means) was used to test the standard data set, and the Food mart data set was also used to compare the RFM-SS-means algorithm with the traditional K-means algorithm and the two-steps algorithm Class effect. From the experimental results, it can be seen that the CH coefficient of RFM-SS-means is the largest, and the clustering effect is good without prior determination of K value and initial clustering center.

Key words: customer classification; semi-supervised clustering; K-means; RFM model

1　引言

随着信息科技的不断进步, 企业经验与管理的理念不停在发生变化, 企业营销经历了从“以产品为中心”到“以服务为中心”, 再从“以服务为中心”到“以客户为

① 基金项目: 四川省高等教育改革项目 ([2014]156551)
收稿时间: 2017-02-21; 修改时间: 2017-03-23; 采用时间: 2017-03-29

中心”不断变迁[1,2]. 客户关系管理 (Customer Relationship Management, CRM) 成为最主流的“以客户为中心”的管理理念之一, 适应企业发展的需求, 近年来备受关注. 客户分类作为 CRM 的重要管理工具, 它是企业进行市场营销的重要依据[3-6]. 它对客户群体进行划分, 方便确定高价值顾客, 从而帮助企业重点关注高价值顾客群体[5]. 在众多的 CRM 的分析模式中, RFM 模型备受瞩目, 被各个不同领域广泛使用. RFM 模型的参考变量是消费者消费行为的数据, 不涉及客户私人隐私且容易得到, 且此模型比较简单便捷, 实践起来也比较有效. RFM 模型是根据客户购买行为进行客户价值评估的模型, 其通过 R(recency)、F(frequency)、M(monetary) 三个变量, 即最后一次购物时间 R、某期间内的购买次数 F 以及某期间内客户花费总金额 M 来描述客户行为价值[7]. 传统的基于 RFM 模型的分类具有划分群组过多、消费的次数 (F) 和总金额 (M) 二者间有着共线性以及不同簇之间差距不大等缺点, 近年来客户分类方法主要有 K-means、Two-step 及 Kohonen 神经网络算法等[8-12], 这些算法往往自身存在不足, 例如对初始聚类中心以及聚类数目的选取比较敏感、对离群点敏感、不适用于大数据等[9]. 其中, K-means 是基于划分的算法, 其运行速率较快且具有很高的运行效率, 它的复杂程度为 O(nkt), 其中, n 为数据库中实体数, t 是迭代次数, k 为簇的数目, 一般来说, $k<<n$, $t<<n$. K-means 算法对大量数据集的处理, 依然有着较高的效率和伸缩性, 但划分数目 K 是 K-means 算法必要的输入的变量, 它对运算结果影响较大且很难确定, 随机确立的初始聚类中心也会导致大相径庭的甚至错误的聚类结果[13]. 同时, 传统的聚类算法由于未考虑客户数据结构的特征, 从而导致聚类过程存在盲目性. 本文采用半监督聚类算法对客户分类展开研究, 旨在减少或弥补上述客户分类算法的不足.

2 基于 RFM 模型的半监督聚类算法

半监督聚类算法是在聚类算法的基础上进行的, 它的核心在于标记信息的处理以及标记信息与聚类算法的结合, 即如何利用标记信息来控制或引导聚类的过程或者调整聚类结果, 使得聚类结果最大程度上按照用户的意愿进行.

2.1 RFM 模型客户数据集先验结构化信息

半监督聚类算法在实际应用中标记样本的方法主要有 Wagstaff 提出的 Must-link 和 Cannot-link 约束方法[14]. 假定给定数据集 D, 已知 P、Q∈D, 如果认为 P、Q 应当在同一类中, 则 Must-link(P, Q)=True, 如果认为两者不应当在同一类中, 则 Cannot-link(P, Q)=True. Must-link 和 Cannot-link 具有如下性质.

性质 1. 对称性. 设 D 为数据集, P、Q∈D, 则如下两式成立:

$$\text{Must}-\text{link}(P,Q)=\text{True}\Leftrightarrow\text{Must}-\text{link}(Q,P)=\text{True} \quad (1)$$

$$\text{Cannot}-\text{link}(P,Q)=\text{True}\Leftrightarrow\text{Cannot}-\text{link}(Q,P)=\text{True} \quad (2)$$

性质 2. 有限的传递性. 设 D 为数据集, P、Q、R∈D, 则下式成立:

$$\text{Must}-\text{link}(P,Q)=\text{True},\ \text{Must}-\text{link}(Q,R)=\text{True}\Rightarrow\text{Must}-\text{link}(R,P)=\text{True} \quad (3)$$

Kotler 作为客户营销战略的倡导者认为关系营销的重心要放在如何和最有价值的少部分顾客建立长期并为公司带来利润的关系[15], 而 Morgan 和 Hunt 更明确指出顾客价值已经成为顾客关系营销的一个中心和重点[16], Wyner 也指出企业 80% 的利润是来自于 20% 的顾客, 而其余 20% 的利润, 却花了公司 80% 的营销费用[17].

在进行 RFM 模型的客户数据结构分析时, 我们可以发现该模型的三个评价 R、F、M 指标存在多重共线性, 例如一个消费额大的客户通常消费频率较高, 同时近期内有购买记录的可能性也较大. 我们参考传统的 RFM 模型划分方式, 将 F、M 两个影响因子从大到小排序, 分别各选取前 20% 的两组客户数据集, 然后再将 R 从近到远排序, 选择前 20% 的客户数据集, 这样我们就得到三组客户数据, 然后筛选出在三组客户群组中都存在的客户, 组成一个客户集 D_1. 同理, 将 F、M 两个影响因子从小到大排序, 各选取前 20% 的两组客户数据, 然后再将 R 从远到近排序, 选择前 20% 的客户数据, 这样我们就得到三组客户数据, 然后筛选出在三组客户群组中都存在的客户, 组成一个客户集合 D_2.

假设 P_1 和 Q_1 是属于 D_1 中的任意两个数据, P_2 和 Q_2 是属于 D_2 中的任意两个数据, 即 P_1、$Q_1\in D_1$, P_2、$Q_2\in D_2$, 则 Must-link(P_1, Q_1)=True, Must-link(P_2, Q_2)=True, Cannot-link(P_1, P_2)=True, Cannot-link(P_1, Q_2)=True, Cannot-link(P_2, Q_1)=True, Cannot-link(Q_1, Q_2)=True. 通过计算可以得出集合 D_1 和 D_2 的 Must-link 集合 M 和 Cannot-link 集合 C.

2.2 成对约束

利用标记信息来控制或引导聚类的过程或者调整聚类结果的半监督聚类算法能使得聚类结果最大程度上按照用户的意愿进行, 提高其效率和精度. HMRF-KMeans[18]是近年来提出的一种半监督聚类算法, 它基于马尔科夫随机场 (Hidden Markov Random Field, HMRF) 成对约束, 目标函数为:

$$J(\pi)=\underbrace{\sum_{c=1}^{k}\sum_{x_i\in\pi_c}\|x_i-m_c\|^2}_{\text{标准的}K\text{-means 目标函数}}+\underbrace{\sum_{\substack{x_i,x_j\in M\\ s.t.\,l_i\neq l_j}}w_{ij}+\sum_{\substack{x_i,x_j\in C\\ s.t.\,l_i=l_j}}\bar{w}_{ij}}_{\text{约束惩罚}} \quad (4)$$

其中, M, C 分别是给定的 Must-link 集合与 Cannot-link 集合; m_c 为聚类中心, w_{ij}, $\overline{w_{ij}}$ 分别为违反 Must-link 和 Cannot-link 约束规则的惩罚权重. 在应用成对约束的半监督聚类方法中, 满足 $l_i=l_j$ 的约束集合被称为 Must-link 集合. 满足 $l_i\neq l_j$ 的约束集合为称为 Cannot-link 集合. HMRF-KMeans 算法使用欧式距离或余弦相似度计算权重.

2.3 自适应确定 K 值

基于传统 K-means 算法自适应确定 K 值的主要思想是: 首先使用标记出的两组数据集, 将该两组数据集的中值作为两个初始聚类中心; 将剩余数据对象按照与这两个初始聚类中心的距离来划分类; 然后将这些距离中距离聚类中心最远的点作为新的聚类中心, 对所有的数据进行二次划分; 如此循环反复, 直到满足条件算法结束. 基于传统 K-means 算法自适应生成 K 值的算法流程如图 1 所示.

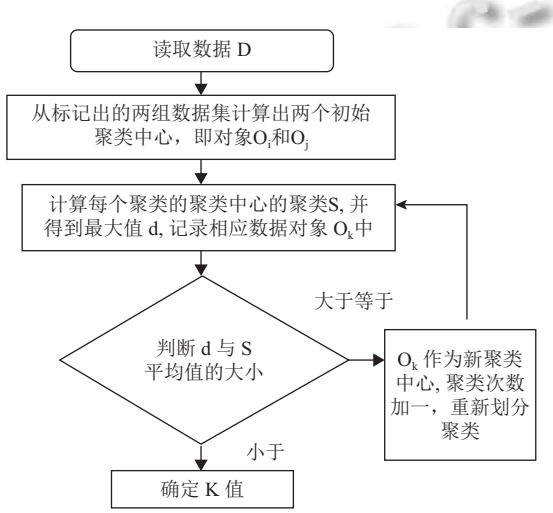


图 1 自适应 K 值方法流程图

2.4 改进算法的实现

通过对 RFM 模型客户数据集先验结构化信息的处理, 利用 HMRF-KMeans 成对约束思想和自适应确定传统 K-means 算法 K 值的方法, 本文提出一种基于 RFM 模型的半监督聚类算法 (Based on RFM model semi supervised K-means, RFM-SS-means), 图 2 是改进后的算法总体流程图, 其中虚线框内标明的是较传统 K-means 算法改进的地方.

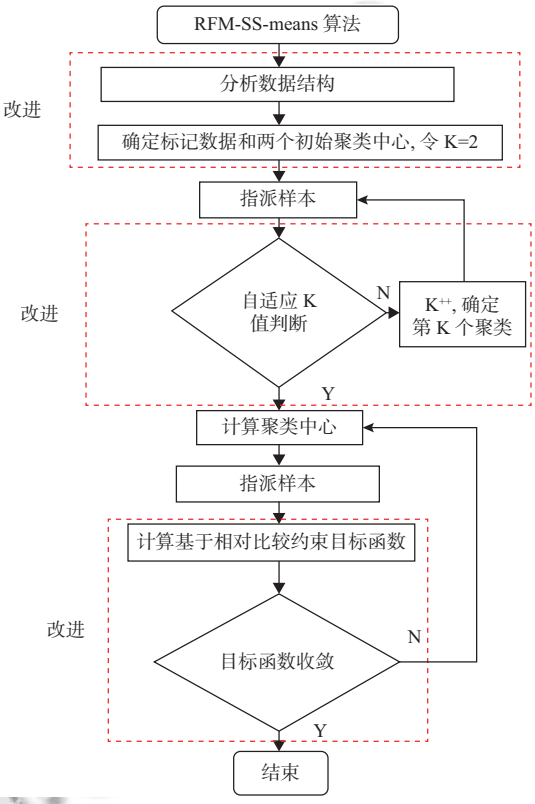


图 2 改进的半监督聚类算法 (RFM-SS-means)

RFM-SS-means 算法挖掘了 RFM 数据集中本身潜藏的先验信息, 使得聚类算法能够获取更多的启发式信息, 从而减少了搜索过程的盲目性, 使用标记数据选取初始聚类中心, 避免了随机获取初始聚类中心导致的聚类结果的不稳定性, 使用自适应确定 K 值的方法确定聚类数目, 解决了聚类之前必须事先确立 K 值的不足. 基于 HMRF-KMeans 成对约束, 在传统的目标函数上引入约束惩罚项和约束奖励项, 实现了对聚类引导和聚类结果的调整.

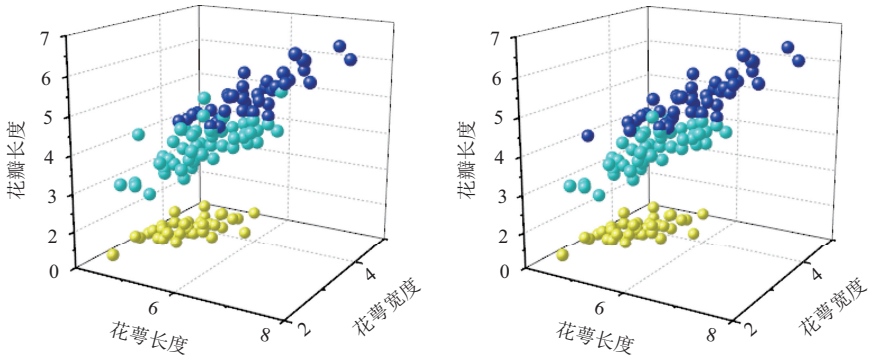
3 实验结果

3.1 标准数据集测试

UCI 数据库 (http://archive.ics.uci.edu/ml/) 是一个

国际上专业进行机器学习、数据挖掘算法测试的标准数据库. 选择 UCI 数据库中的 Iris 和 Wine 这两个标准数据集作为测试数据集, 测试 RFM-SS-means 算法聚类的准确性.

以数据集 Iris 为例, 根据分类结果, 选择 Iris 每个样本的三个属性 (花萼长度、花萼宽度和花瓣长度) 画出三维原始数据的分类效果图, 实现可视化的目的, 聚类效果如图 3 所示.



(a) RFM-SS-means (b) 标准 Iris

图 3 RFM-SS-means 算法与标准 Iris 的聚类效果

为了更加准确的检验聚类算法的性能, 采用外部有效性评价指标 FMI 指标将聚类结果与“真实”聚类结果进行比较. 随机选择 Iris 和 Wine 数据集中的 15、30、45、60、75、90、105、130、145 个数据对象共 18 组数据, 抽取数据对象时从每个类抽取同等数量的数据对象, 进行准确性验证, 得到结果如表 1 所示.

表 1 RFM-SS-means 算法的外部有效性评价结果

数据总量(个)	Iris的FMI值	Wine上FMI值
15	1	0.734
30	1	0.527
45	0.856	0.802
60	0.874	0.732
75	0.920	0.710
90	0.905	0.756
105	0.858	0.687
130	0.931	0.771
145	0.912	0.750
平均值	0.917	0.719

FMI 指标结果区间为[0, 1], 值越大表示聚类效果越好, 当 FMI 指标为 1 时, 代表分类结果与“真实”分类情况完全一致. 通过表 1 可以看出, RFM-SS-means 算法在 Iris 数据集上的平均 FMI 指标达到了 0.917, 在 Wine 数据集上的 FMI 指标达到了 0.719, 具有较高的聚类准确性.

3.2 Food Mart 数据库测试

考虑到 UCI 标准数据集并不具备 RFM 客户数据集的结构特征, 难以发挥标记数据的作用, 因此使用 Food Mart 数据集对改进算法进行测试. Food Mart 数据库是根据某食品超市在 1998 年的售货记录得到的, 该数据库 RFM 模型的基本特征. 该数据库包含 23 张原始表, 主要的数据表为 sales_fact_1998(1998 年的商品交易数据), customer(顾客属性信息表, time_data(商品销售时间信息) 等. 顾客交易资料中关键字段解释如表 2 所示.

表 2 Food Mart 数据库字段字典

字段名	字段解释
Product_id	商品ID
Time_id	商品出售时间ID
Customer_id	客户ID
Promotion_id	促销信息ID
Store_id	商店ID
Store_sales	商店出售信息
Store_cost	顾客购买商品单价
Unit_sales	顾客购买商品总数

通过对 Food Mart 数据库中 321 个客户的相关客户信息、商品销售信息、商品销售时间等数据表进行处理, 可以得到 RFM 模型的三个维度的输入变量: 客户总消费 (M)、消费次数 (F) 和最近消费时间间隔 (R). 运用本文前面提到的改进方法, 对 Food Mart 数据库客户数据的先验结构化信息进行分析, 可以得到两组消费行为迥异的用户族群, 将这两组客户数据作为

先验类别标记, 进而可以得到 RFM-SS-means 算法的两个初始聚类中心.

分别应用 K-means、two-step 和 RFM-SS-means 算法对客户群体进行分类, 根据三种聚类方法的分类结果, 作出三维原始数据的客户分类效果图, 如图 4 所示.

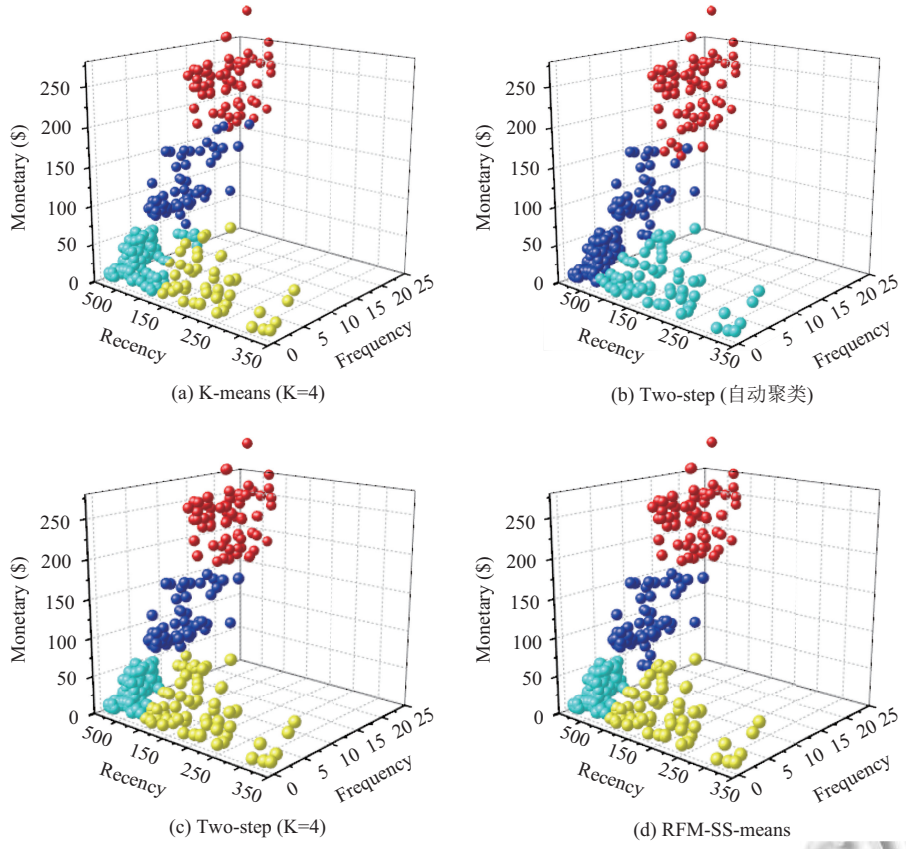


图 4 三种聚类算法聚类效果图

为了准确的检验聚类算法的性能, 采用内部检验法对三种方法的聚类结果进行效果评价, 即分别计算出不同聚类结果的 CH 系数, 得到如表 3 所示结果.

CH 系数是基于对观测数据的组内距离差矩阵与组间距离差矩阵的测度, CH 系数的值越大, 则表明分类效果就越好. 从表 3 可以得到 Two-step、K-means 与 RFM-SS-means 算法的 CH 系数分别为 8.52, 9.14, 9.31, 可以看出本文提出的半监督聚类算法聚类效果最优.

表 3 不同聚类算法聚类效果的内部检验法评价 (CH 系数)

聚类方法	K-means (K=4)				Two-step (K=4)				RFM-SS-means			
聚类类别	红	深蓝	浅蓝	黄	红	深蓝	浅蓝	黄	红	深蓝	浅蓝	黄
均值(R)	9.3	24.5	33.0	188.2	9.30	24.7	21.4	21.1	9.30	25.9	21.4	164.1
均值(F)	22.3	11.4	3.9	3.8	22.3	10.9	3.9	4.0	22.3	11.3	3.9	3.7
均值(M)	216.9	93.0	23.4	23．7	216.9	93.1	23.0	23.6	216.9	91.7	23.0	23.4
类内距离平方和	1399055				1398416				1378465			
类间距离平方和	157609				168921				169744			
CH系数	8.52				9.14				9.31			

4 总结

本文提出了一种基于 RFM 模型和 K-means 算法的半监督聚类算法 (RFM-SS-means). 通过对客户数据集中本身潜藏的先验结构化信息进行研究, 标记出两

组客户数据作为类别标记, 进而计算出两个初始聚类中心, 同时采用自适应方法自动确定 K 值和初始聚类中心, 解决了传统算法中 K-means 算法的 K 值和初始聚类中心难以确定的不足. 引入 Must-link 和 Cannot-link 两种约束将类别标记转换为成对约束信息, 基于 HMRF-KMeans 成对约束, 在传统的目标函数上引入约束惩罚项, 从而实现对聚类引导和聚类结果的调整, 提高了聚类实现的效率. 本文提出的 RFM-SS-means 算法可用于企业对客户群体的分类, 有利于企业针对不同客户群体开展精准的营销活动, 提高企业的营业收入和利润.

参考文献

1 王海东. 客户关系管理实施要点. 中国电力企业管理, 2007, (11): 58. [doi: 10.3969/j.issn.1007-3361.2007.11.031]

2 何荣勤. CRM 原理•设计•实践. 北京: 电子工业出版社, 2003: 3–12.

3 Gloy BA, Akridge JT, Preckel PV. Customer lifetime value: An application in the rural petroleum market. Agribusiness, 1997, 13(3): 335–347. [doi: 10.1002/(ISSN)1520-6297]

4 马椿荣. 消费者价值研究理论综述. 商业时代, 2014, (10): 60–61. [doi: 10.3969/j.issn.1002-5863.2014.10.025]

5 Hughes AM. Strategic database marketing: The masterplan for starting and managing a profitable, customer-based marketing program. Irwin Professional, 1994.

6 Duboff RS. Marketing to maximize profitability. The Journal of Business Strategy, 1992, 13(6): 10–13. [doi: 10.1108/eb039521]

7 Cheng CH, Chen YS. Classifying the segmentation of customer value via RFM model and RS theory. Expert Systems with Application, 2009, 36(3): 4176–4184. [doi: 10.1016/j.eswa.2008.04.003]

8 徐翔斌, 王佳强, 涂欢, 等. 基于改进 RFM 模型的电子商务客户细分. 计算机应用, 2012, 32(5): 1439–1442.

9 闫相斌, 李一军, 叶强. 基于购买行为的客户分类方法研究. 计算机集成制造系统, 2015, 11(12): 1769–1774.

10 郑茜茜. 基于数据挖掘的客户细分研究[硕士学位论文]. 重庆: 重庆交通大学, 2013: 8–10.

11 刘芝怡, 陈功. 基于改进 K-means 算法的 RFAT 客户细分研究. 北京理工大学学报, 2014, 38(4): 531–536.

12 潘玲玲, 张育平, 徐涛. 核 DBSCAN 算法在民航客户细分中的应用. 计算机工程, 2012, 38(10): 70–73. [doi: 10.3969/j.issn.1000-3428.2012.10.020]

13 张建萍, 刘希玉. 基于聚类分析的 K-means 算法研究及应用. 计算机应用研究, 2007, 24(5): 166–168.

14 Wagstaff K, Cardie C. Clustering with instance-level constraints. Proc. of the 17th International Conference on Machine Learning. San Francisco, CA, USA. 2000. 1103–1110.

15 Kotler P. Marketing management.Upper Saddle River, NJ: Prentice Hall, 2000.

16 Morgan RM, Hunt SD. The commitment-trust theory of relationship marketing. Journal of Marketing, 1994, 58(3): 20–38. [doi: 10.2307/1252308]

17 Wyner GA. Customer valuation: Linking behavior and economics. Marketing Research, 1996, 8(2): 36–38.

18 Basu S, Bilenko M, Mooney RJ. A probabilistic framework for semi-supervised clustering. Proc. of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA. 2004. 59–68.