支持交叉营销的金融产品客户数据挖掘①

黄　洪　洪　毅 (浙江工业大学 软件学院 浙江 杭州 310023)

摘　要： 首先比较了 DBSCAN,CLIQUE,CLARANS,K-means 和 X-means 等聚类算法,接着选用 X-means 聚类算法建立了金融产品客户细分模型，然后结合关联强度分析，设计了支持交叉营销的金融产品客户数据挖掘系统，并给出了一个系统使用示例。

关键词： 交叉营销；X-means 聚类算法；客户细分；金融产品营销

Customer Data Mining for Supporting Cross-Marketing of Financial Products

HUANG Hong, HONG Yi

(Software College, Zhejiang University of Technology, Hangzhou 310023, China)

Abstract: This paper makes a comparison between Clustering algorithms such as DBSCAN, CLIQUE, CLARANS, K-means and X-means. The X-means clustering algorithm is selected to establish a customer segmentation model for financial products marketing. Based on relational analysis of financial products, a financial products customer data mining application system is designed to support the cross-marketing of financial products. In the end, a use case is given to illustrate the application of the system.

Keywords: cross-marketing; X-means clustering algorithm; customer segmentation; financial product marketing

交叉销售是指通过发现现有顾客的多种需求，并进一步满足需求来实现向该顾客销售多种相关的服务或产品的销售方式。交叉营销就是促成交叉销售的各种策略和方法的总和。随着市场竞争越来越激烈，谁留住客户，谁就是最后的赢家。研究表明，一个企业将一种产品或服务推销给一个老客户的成本远低于吸收一个新客户的成本，而一个客户如果购买了一个企业四个以上的产品或服务后，其客户的流失率几乎是零[1]。综上所述，企业通过有效地实施交叉营销，可以降低营销成本、增加利润、提高客户的忠诚度。

在金融企业开展交叉营销具有得天独厚的有利条件，因为消费者在购买金融产品或服务时必须提交真实的个人资料，这些资料为进一步分析顾客的需求和潜在再消费能力提供了基础数据[2]。目前已经有企业在实行交叉营销，但当前的交叉营销大都是在缺乏对客户和产品的科学分析的情况下开展的，没有针对性，因而效率很低。

针对以上分析，本课题的研究者认为，要进行有效的交叉营销，必须以客户细分作为基础，因为只有充分了解不同客户的特征，才能有针对性的进行基于客户个体的交叉营销，从而降低营销成本、提高营销效率。数据挖掘是从已有数据中发现未知规律的有力工具，数据挖掘的聚类算法非常适合用于对客户进行细分，但聚类算法的选择直接影响到对客户数据进行分析的效率和科学性。

因此本文主要工作是：首先比较了多种聚类算法，接着选用 X-means 聚类算法建立了金融产品客户细分模型，然后给出了产品关联强度分析方法，最后结合客户细分模型和产品关联分析，设计了支持交叉营销的金融产品客户数据挖掘系统。

1　金融产品客户细分聚类算法的选择

要实施好交叉营销，关键是要进行有效的客户分析，即客户细分。数据挖掘技术中，聚类算法已成为实现客户细分的目标的最重要的工具，比较常用的聚类算法有 DBSCAN，CLIQUE，CLARANS，K-means

① 收稿时间:2009-09-07;收到修改稿时间:2009-10-21

和 X-means 等。本节首先比较这些算法，并结合金融产品客户数据的海量以及复杂的特性，给出选用 X-means 聚类算法的主要理由。

(1) DBSCAN 算法：对于一个类中的每个对象，在其给定半径的领域中包含的对象不能少于某一给定的最小数目。该算法遇到数据量非常大时，由于缺乏对数据的预处理而直接对整个数据集进行操作，消耗大量的内存和 I/O。同时，用户需要事先设定聚类对象的半径和最小包含对象数，具有一定的难度。

(2) CLIQUE 算法：该算法是基于密度和网格的聚类方法，根据规则求出各个子空间的聚类单元。但是由于方法过于简化，同时缺少对原始数据特性的分析，硬性的网格划分，增加了计算的复杂度，降低了聚类结果的精确性。

(3) CLARANS 算法：是一种分割聚类方法。该算法是在 Clara 算法的基础上提出的，主要弥补了 Clara 算法的不足——效率取决于采样的大小，难以得到最佳结果。但是该算法也存在弊端：在每一次循环的过程中的采样都不同，而且需要人为地限定循环的次数。这对大数据量而言，无论时间复杂度还是空间复杂度都相当大。

(4) K-means 算法：是一种分割式聚类方法，其主要目标是要在大量高维度的数据点中找出具有代表性的点。但是事先需要指定聚类数 K，而要事先指定一个合理的聚类数是非常困难的。

(5) X-means 算法：能够在一个指定的聚类数范围[R1，R2]内找到一个最优的聚类数 R，实现聚类划分，使得聚类分析的结果更具合理性，此外在对海量数据在执行效率、异常点检测能力等方面较其他算法有明显的优势[3-5]。

通过上述比较，我们可以看出 X-means 算法更加适合处理包含较多噪声的复杂海量金融产品客户数据，因此本文选用了该算法。

X-means 聚类算法的基本流程是：从下限值 R1 开始，运行 K-means 算法(分裂过程：将 n 个数据对象划分为 K 个聚类以便使得所获得的聚类满足：同一聚类中的对象相似度较高；而不同聚类中的对象相似度较小，其中聚类相似度是利用各聚类中对象的均值所获得一个“中心对象”来进行计算的)，直到满足算法结束条件时终止(该条件指参数 R 达到上限 R2 或者没有聚类中心需要重新分裂)。最后根据各聚类本身尽可能的紧凑，而各聚类之间尽可能的分开的原则，从范围[R1，R2]内找到一个最优的聚类数 R，其中 K-means 算法的流程如下：(1)从 N 个数据对象中任意选择 R1 个对象作为初始聚类中心；(2)循环(3)到(4)直到每个聚类不再发生变化为止；(3)根据每个聚类对象的均值(中心对象)，计算每个对象与这些中心对象的距离，并根据最小距离重新对相应对象进行划分；(4)重新计算每个(有变化)聚类的均值(中心对象)。

其中涉及到的不同个体对象与某个中心对象间的距离计算实现过程如下：逐个将对象 x_i(i=1，2，…，n)按欧式距离分配给距离最近的一个聚类中心 c_j，1<j<R，$|x_i-c_j|=\min_{1<j<R}\sqrt{\sum_{l=1}^{m}(x_{il}-c_{jl})^2}$，(其中 m 是数据属性的个数)。java 语言的语法形式表示为：

```
int length = object.length;//个体对象长度
for (int i=0; i<length; i++){//遍历对象
double distance = object[i]-center;//计算距离
double square+= distance * distance;}
Math.sqrt(square);}
```

[6,7] //计算平方差

其中 object 代表聚类对象数组，object[i]是对象数组中的第 i 个个体对象，center 代表中心对象。

2 金融产品客户细分模型的建立和产品关联度分析

本文基于 X-means 聚类算法从性别、年龄、职业、年收入等多要素入手建立客户细分模型，具体步骤如下：

第一步：客户历史交易数据的收集和筛选分析，并根据领域专家的经验确定聚类参数 R 的搜索范围[R1,R2]。

第二步：在寻得最优的聚类数 R 后，分析每个聚类号所对应的各项特征数据，构建客户细分模型。本文设定所关注的特征项主要是客户的比重、男女比例、年龄跨度、职业、年收入、交易跨度、交易频率等。最终建立具体模型如表 1 所示：

表 1　客户细分模型

聚类号	比重	男女比例	年龄跨度	职业	年收入	交易跨度	频率
1	N_1	B_1:G_1	A_0,A_1	P_1	I_1	T_0,T_1	F_1
2	N_2	B_2:G_2	A_1,A_2	P_2	I_2	T_1,T_2	F_2
…	…	…	…	…	…	…	…
R	N_R	B_R:G_R	A_{R-1},A_R	P_R	I_R	T_{R-1},T_R	F_R

根据客户细分结果(对应聚类号)，可以判断某类客户往往会倾向购买某类金融产品，因此对于符合某类

特征的客户，可以采取相应的营销策略。

第三步：产品关联强度的分析。统计所有购买了产品 i 的所有客户数(设为 K_i)，再统计这 K_i 个客户中有多少个购买了产品 j(设为 K_j)，则 $R_{ij}=K_j/K_i$ 就是产品 i 与产品 j 的关联强度，R_{ij} 说明了购买了产品 i 的客户有多大的概率会购买产品 j。这里采用矩阵分析产品关联度：具体见表 2 所示。

表 2　产品关联强度表

关联度	产品 1	产品 2	…	产品 N
产品 1	1	R_{12}	…	R_{1N}
产品 2	R_{21}	1	…	R_{2N}
…	…	…	1	…
产品 N	R_{N1}	R_{N2}	…	1

第四步：根据客户聚类分析和产品的关联分析，进行客户特征归纳和描述，形成客户对号入座的描述表(见表 3)。

表 3　客户群描述表

聚类号	客户特征描述	交叉营销建议
1	C_1	D_1
2	C_2	D_2
…	…	…
R	C_R	D_R

根据产品的关联强度可以判断该客户购买产品 j 后，往往倾向于购买产品 k，那么在客户购买产品 j 后就可以向其进行产品 k 的交叉营销，大大提高了组合销售或者交叉销售的成功率。

第五步：交叉营销实施。根据客户的基本信息(姓名、年龄、年收入、职业等)和客户细分结果，判断该客户所属分类，然后根据该类客户的特征判断其消费的偏向，再根据其已经购买的产品以及产品间的关联关系提出相应的交叉营销建议。

3　支持交叉营销的金融产品客户数据挖掘系统的设计

(1) 系统概述

本系统主要立足交叉营销意义——预测金融消费者商品购买行为，同时致力于最大程度地识别隐含的产品组合消费潜力，有助于产品定位和交叉销售。本系统的基本流程可以概述为以下三步：①通过 ETL(数据抽取、转换和加载)技术，对海量历史记录进行预处理，清除其中无用的噪声数据，对数据进行适当的转换。②选用 X-means 聚类算法，对客户数据按特征进行聚类分析，建立客户细分模型，得出客户细分描述，并形成知识库。③对输入的客户信息，根据知识库进行分析和决策，给出交叉营销建议。

(2) 系统基本结构

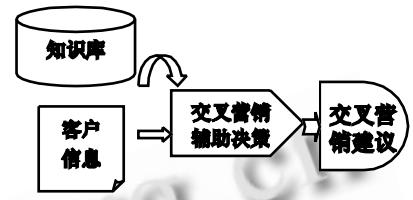

图 1　系统基本结构图

(3) 系统基本设计流程

第一步：聚类分析，客户细分模型建立的简单过程。如下图 2 所示，客户细分模型主要建立在客户信息表和产品消费表的基础上，形成分类表。

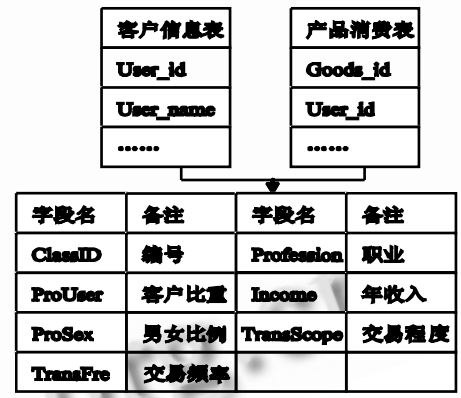

图 2　客户细分模型建立流程图

第二步：产品关联强度建立的简单流程。

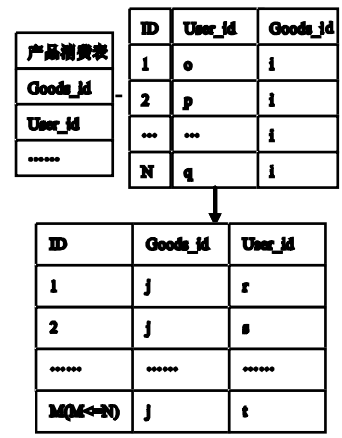

图 3　产品关联强度建立流程图

计算方法具体见本文第 2 部分的第三步的“产品关联强度的分析”。

第三步：结合客户细分模型和产品关联强度，建立客户群描述表，见表 4(客户群描述表)。

表 4 客户群描述表

字段	备注	字段	备注
ID	聚类号	Suggestion	营销建议
Character	特征描述		

(4) 系统应用示例

系统应用场景：在信用卡的月账单上发布针对个体客户的交叉营销广告。

系统应用前提：已通过本文介绍的方法获得相关知识，形成了辅助决策知识库。

系统应用过程：①系统获得客户的基本信息，如身份证号、姓名、年龄、年收入、职业等；②系统搜寻该客户已有交易信息，包括其购买的相关金融产品类型、金额，使用信用卡进行支付情况等；③系统根据客户数据和客户细分模型，确定该客户聚类归属，根据该类的客户特征，获得针对该客户的营销建议，并根据其已购买的产品和产品的关联关系，形成针对该客户的交叉营销方案；④随信用卡账单生成营销广告，反映交叉营销方案，实现个性化交叉营销。

4 结语

本文从交叉营销的时代背景出发，通过对 DBSCAN，CLIQUE，CLARANS，K-means 和 X-means 等聚类算法的比较，选用 X-means 算法建立客户细分模型，并详细阐述了模型建立的步骤。同时，根据模型，结合关联分析，在理论上设计了支持交叉营销的金融产品客户数据挖掘应用系统，并给出了一个模拟应用示例，显示了数据挖掘系统在金融业交叉营销中良好的应用前景。

参考文献

1 英国《经济学人》杂志社编.王昕璐译.营销智典.北京:中信出版社, 2005,11.12－26.

2 霍金斯,贝斯特,科尼.消费者行为学.北京:机械工业出版社, 2002. 20－28.

3 Pelleg D, Moore A. X-means: Extending K-means with Efficient Estimation of the Number of Clusters. School of Computer Science, Carnegie Mellon University, Pittsburgh, PA15213 USA.

4 Torra V. Modeling Decisions for Artificial Intelligence. Springer, 2005,9.140－156.

5 Myatt GJ. Making sense of data. Wiley-Interscience, 2006,11.32－35.

6 Holweg M, Pil FK. Successful Build-to-order Strategies Start with the Customer. MIT Sloan Management Review, 2001,(43):74－83.

7 Quinlan JR. C4.5.-Programs for machine learning. San Mateo: Morgan Kaufmann Publisher, 1993. 170－247.