

# 改进 K-means 算法实现移动通信行为特征分析

何云,李辉,姚能坚,赵榕生

(广州军区空军指挥自动化工作站,广东 广州 510071)

**摘要:** K-means 算法被广泛用于客户细分聚类应用研究,客户细分对移动通信行业具有重要的商业价值。但变量的量纲、维度、聚类数、初始聚点等参数的计算是影响 K-means 算法聚类应用效果的重要因子。在基于 K-means 算法移动通信行为特征分析系统的实现过程中,分别从特征维度选择、变量量纲统一、聚类数 K 值与初始聚点的确定等四个方面改进算法的上述影响参数的计算方法,并利用经验加权的方式使算法与主观经验结合。研究结果表明改进 K-means 算法对移动通信特征分析客户聚类有效。

**关键词:** 客户细分; K-means; 影响因子

**中图分类号:** TP301.6

**文献标识码:** A

**文章编号:** 1673-629X(2011)06-0063-03

## Application of Improved K-Means Algorithm in Mobile Communication Behavioral Characteristic Analysis

HE Yun, LI Hui, YAO Neng-jian, ZHAO Rong-sheng

(Command Automation Office, Guangzhou Military Region Air Force, Guangzhou 510071, China)

**Abstract:** K-means algorithm is widely used to customer segmentation clustering application research, customer segmentation of mobile communications has important commercial value. But dimension unit, dimension of variable, cluster numbers, initial centroids, etc. calculation of these parameters is important factor of influencing K-means algorithm cluster application result. Based on K-means algorithm mobile communication behavior characteristic analysis process of implementing, respectively from the characteristic dimensions selection, variable dimension unit unity, cluster number K value and initial centroids determination four aspects, improve the determination of the above algorithm affects parameters calculation method, utilize experience weighting way to make algorithm bind with subjective experience. The result of study indicates that according to behavioral characteristic analysis improving K-means algorithm will subdivide the cluster to the mobile communication customer effectively.

**Key words:** customer segmentation; K-means; influencing factor

## 0 引言

客户细分对移动通信行业具有重要的商业价值,它在获取潜在客户、减少客户流失、降低服务成本、提升客户价值、提高满意度、制定精确化营销策略等方面具有重要意义<sup>[1-3]</sup>。建立移动通信行为特征分析的目的就是要将 AC 接口信令解码数据、BOSS 计费数据、GPRS 上网数据等通信行为特征数据用于辅助营销决策。通过对客户接打电话、收发短信、上网以及开关机等通信行为进行数据挖掘,将隐藏在这些数据中的尚未被发现的知识提取出来,使企业更清楚地了解用户通信行为差异化特征,以便市场部门可以对不同的用户制定不同的营销策略,并且在宣传推广、新业务交叉

营销中发挥作用,进一步巩固和发展与客户的关系。K-means 算法被广泛用于客户细分聚类应用研究,但 K-means 算法的效果受变量的量纲、维度、聚类数、初始聚点等因子的影响很大<sup>[4-6]</sup>,大量应用研究表明上述影响因子与具体的案例和主观经验相关联<sup>[7-9]</sup>。

文中试图通过基于移动通信行为特征分析的客户聚类实例改进 K-means 算法的上述影响参数的计算方法,并利用经验加权的方式使算法与主观经验结合。研究结果表明改进 K-means 算法对移动通信行为特征分析客户聚类有效。

## 1 K-means 模型介绍

### 1.1 K-means 算法实现

BOSS 数据、AC 接口信令解码数据、GPRS 数据通过采集、抽取、转换、加载到数据仓库等步骤形成通信行为特征数据,通信行为特征包括语音行为、短信行为、上网行为等多种行为特征。每个记录包含实体标

收稿日期:2010-10-26;修回日期:2011-01-27

基金项目:广空预研项目(GK2009BE0102)

作者简介:何云(1974-),男,湖北天门人,硕士,工程师,CCF 会员,研究方向为通信技术、网络安全。

识和多个特征数据,通过对记录特征数据的聚类分析将数据表中的实体标识分为不同的细分类,达到依据通信行为特征对客户进行细分的目的。

聚类(clustering)是将一批数据依据它们的相似属性归类,使人们能够对数据进行概括性的理解。聚类分析是根据事物本身潜在的特性研究对象分类的方法。通过聚类把一个数据集中的个体(对象)按照相似性归约成若干类别,使其“物以类聚”。聚类分析的原则是使同一类别中的对象之间具有尽可能大的相似性,而不同类别中的对象之间具有尽可能大的差异性。与分类分析不同的是,聚类结果主要基于当前所处理的数据,不依赖于预先定义好的类,事先也不知道可分割的类的个数。因此在机器学习中,数据分类被称为监督学习,而数据聚类则称为非监督学习。聚类分析的方法包括基于划分的方法、基于密度的方法、基于层次的方法、基于网格的方法、基于模型的方法等<sup>[9-11]</sup>。

由于移动通信行为特征数据海量的特点,在本实例中聚类分析的算法主要采用 K-means 算法<sup>[12]</sup>。

K-means 算法是简单而有效的统计聚类技术,将样本集根据它们之间的相似程度分为预先制定的  $K$  个组。若定义  $n$  为样本个数,  $K$  为聚类数,则算法的基本步骤可表述如下:

(1) 选择一个  $K$  值,用以确定簇的总数;

(2) 在数据集中任意选择  $K$  个样本,作为初始聚类中心  $c_1, c_2, \dots, c_k$ ;

(3) 依据样本  $\{x_i, i = 1, 2, \dots, n\}$  到聚类中心的欧式距离,将其归入距离它们最近的中心  $c_j$  的簇  $X_j$ , 即若:

$$d(x, c_j) = \min \{d(x, c_i), i = 1, 2, \dots, K\} \quad (1)$$

则  $x \in X_j$ ;

(4) 使用每个类中的样本来计算每个簇新的聚类中心:

$$c_j = \frac{1}{n_j} \sum_{x \in X_j} x \quad (2)$$

其中  $j = 1, 2, \dots, K, n_j$  是类  $X_j$  中的样本数;

记准则函数:

$$J = \sum_{j=1}^K \sum_{x \in X_j} d^2(x, c_j) \quad (3)$$

(5) 如果  $J$  值减少,则转至步骤 3 继续迭代,否则终止。

## 1.2 模型有效度验证

聚类的有效性验证涉及到三个方面的问题<sup>[12]</sup>:一是聚类的质量,二是聚类方法是否适合特定的数据,三是类的最优数量。概括的说,有两种标准来进行聚类评价以及聚类方案的选择:

类内对象之间的紧凑性。类内对象之间的距离应

该尽可能的小,一般用方差来衡量,应该使方差达到最小。

不同类之间的距离尽可能大。衡量标准有:以类间中心点的距离、类间最远点的距离和类间最近点的距离三种。

具体验证方法如下:

对数据集  $\{x_i\}$ ,  $x_i = (x_{i1}, \dots, x_{ip}, \dots, x_{in})$ , 其中  $i = 1, 2, \dots, n$ 。设有  $K$  个类,类  $c_i$  包含了  $n_i$  个样本,其中  $i = 1, 2, \dots, K$ 。数据集的方差为  $\sigma(X) = (\sigma_1, \dots, \sigma_p, \dots, \sigma_s)$ , 其中:

$$\sigma_p = \frac{1}{n} \sum_{j=1}^n (x_{jp} - \bar{x}_p)^2 \quad (4)$$

类  $c_i$  的方差为  $\sigma(c_i) = (\sigma_{i1}, \dots, \sigma_{ip}, \dots, \sigma_{is})$ , 从而得到类内方差:

$$\sigma_{\text{mtra}} = \sum_{i=1}^K \sigma(c_i) \quad (5)$$

则类内紧凑度为:

$$d_{\text{mtra}} = \frac{\|\sigma_{\text{mtra}}\|}{K \|\sigma(X)\|} \quad (6)$$

$d_{\text{mtra}}$  值越小,表明类内距离越小,聚类结果可能越好,但最终结果的好坏还要看类间距离的度量值,即:

$$d_{\text{mter}} = \frac{\sum_{i=1}^K \sum_{j=1}^K \|c_i - c_j\|}{K(K-1)} \quad (7)$$

$d_{\text{mter}}$  值越大,表明类间距离越大,聚类结果可能越好。因此,根据上面类内类间距离的度量值,可得最终聚类结果的评价公式为:

$$V = \frac{d_{\text{mtra}}}{d_{\text{mter}}} \quad (8)$$

显然,  $V$  的值越小,聚类方案越好。

## 2 影响因子的计算方法

在通信行为特征分析实例中,发现 K-means 算法的效果受变量的量纲、维度、聚类数、初始聚点等因子的影响很大。通过实验结合业务意义、细分目的、数据质量等经验总结出变量的量纲、维度、聚类数、初始聚点的计算方法如下:

### 2.1 特征维度选择

总结的特征维度的选择原则有:

(1) 从业务的角度来看,无分析意义的属性不选择;

(2) 离散属性、取值个数较少的及数据质量较差的不选择;

(3) 互相之间可以派生的(相关性特强)不能全部选择;

(4) 选择分量,去掉总量;

(5) 尽可能涵盖所有业务范围。

按照以上原则,现将通信行为基本维度归纳如下(见表 1)。

表 1 通信行为特征基本维度表

| 数据字段   | 描 述         |
|--------|-------------|
| 本机号码   | 通信主体标识      |
| 开机时长   | 合计时长        |
| 主叫次数   | 合计次数        |
| 主叫时长   | 合计时长        |
| 被叫次数   | 合计次数        |
| 被叫时长   | 合计时长        |
| 发送短信量  | 合计次数        |
| 接收短信量  | 合计次数        |
| 长话次数   | 合计次数        |
| 长话时长   | 合计时长        |
| 网间通信量  | 网间语音、短信行为次数 |
| 忙时通话量  | 忙时语音通信次数    |
| 工作日通话量 | 工作日语音通信次数   |
| ...    | ...         |

2.2 变量量纲统一

本实例中遵照选择分量、去掉总量的原则,而特征维变量有的是话务时长(单位有的是小时、有的是 Erl)、有的是统计次数,变量量纲的不一致会导致欧式距离的计算与客观事实不符。采用分量除以总量记百分数的方式可以统一变量量纲。变量的值记为:分量/总量 \* 100 \* 权值,权值大小是依据经验确定,即代表某个特征对客户细分的影响大小。变量量纲统一后的特征维度表见表 2。

表 2 统一后的特征维度表

| 数据字段   | 量 纲                       | 权值范围 |
|--------|---------------------------|------|
| 本机号码   | 实体标识                      |      |
| 开机时长   | 开机时长/统计时长 * 100 * 0.08    | 0.08 |
| 主叫次数   | 主叫次数/总呼叫次数 * 100 * 0.09   | 0.09 |
| 主叫时长   | 主叫时长/通话总时长 * 100 * 0.04   | 0.04 |
| 被叫次数   | 被叫次数/总呼叫次数 * 100 * 0.06   | 0.06 |
| 被叫时长   | 被叫时长/通话总时长 * 100 * 0.05   | 0.05 |
| 发送短信量  | 发送短信次数/总短信次数 * 100 * 0.08 | 0.08 |
| 接收短信量  | 接收短信次数/总短信次数 * 100 * 0.08 | 0.08 |
| 长话次数   | 长话次数/总通话次数 * 100 * 0.12   | 0.12 |
| 长话时长   | 长话时长/通话总时长 * 100 * 0.1    | 0.1  |
| 网间通信量  | 网间通信次数/通信总次数 * 100 * 0.1  | 0.1  |
| 忙时通话量  | 忙时语音通信次数/通信总次数 * 0.08     | 0.08 |
| 工作日通话量 | 工作日语音通信次数/通信总次数 * 0.12    | 0.12 |
| ...    | ...                       | ...  |

采用这种量纲的好处有:统一了变量的单位,欧式距离的计算更客观;选择分量去掉总量后,总量以分母的形式体现到特征中;各个特征间的关联性没有丢失;主观经验以权值的形式体现。

2.3 聚类数 k 与初始聚点的确定

通信行为特征数据量通常在 TB 级别,进行数据挖掘特别耗时,K 值与初始聚点的选择需要利用相对

较小的训练样本数据进行测定。基本思想是给定 K 值的变化范围,使训练样本空间数据集 K-means 聚类效果最优,从而确定 K 值和初始聚点,具体的步骤是:

- (1) 选择数据质量较好的训练样本空间数据集;
- (2) 选取 K 值 = 特征维数;
- (3) 依据 1.1 执行 K-means 聚类模型;
- (4) 依据 1.2 模型验证计算  $V = \frac{d_{mtra}}{d_{mter}}$ ;
- (5) K 值 = K 值 + 1,重复步 3、4 直到 K 达到 2 倍特征维数;
- (6) V 值最小对应的 K 值为最终的 K 值;
- (7) V 值最小对应距离各个类的聚类中心最近的 K 个样本为初始聚点。

3 改进算法的应用效果

在广东某地市移动公司精确化营销项目的实践中,专门对改进算法与原 K-means 算法的客户细分营销效果进行了对比研究。该项目采集某地市移动公司 100 多万移动用户的 AC 口信令数据、上网数据和计费数据,将数据抽取、转换、加载为用户通信行为特征数据集,在数据集上采用不同的模型进行数据挖掘,为用户提供通信行为特征分析客户聚类多维报表服务。

基于通信行为特征分析精确化营销项目体系架构见图 1。

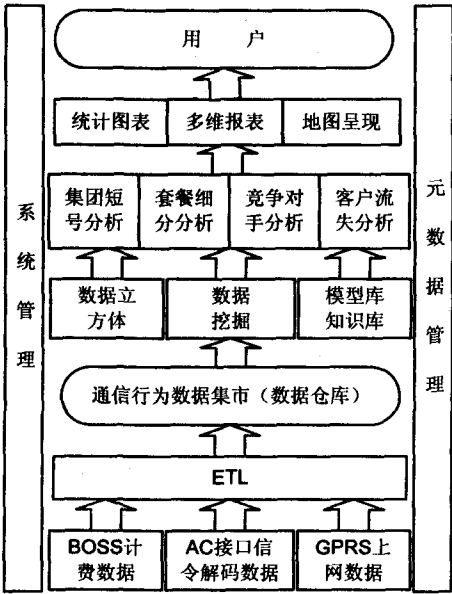


图 1 基于通信行为特征分析精确化营销系统架构

在该项目实施过程中,分别对原 K-means 算法和改进算法进行聚类应用,得到基于通信行为特征分析的客户聚类结果,再依据客户细分进行营销推广。跟踪该项目近一年的营销推广效果,可以看出该改进算

(下转第 69 页)

从图4、图5和图6中可以看到质心轨迹以及足底运动完全符合预期,同时质心偏转较小,说明上身姿态被控制在一个较小的范围内,说明机器人的摆动较小,稳定性较强。

#### 4 结束语

文中给出了一种基于三维线性倒立摆的双足机器人步态规划的算法,通过步行距离、步行周期,以及质心高度等参数,获得质心的轨迹。本方法适用于双足机器人的大多数步行方式,特别是对于角度不大的转向运动具有极强的稳定性。本算法忽略了力矩的影响,因为从实际试验来看在转向角度不大的情况下力矩对于ZMP点的影响不大。

#### 参考文献:

- [1] 蔡自兴. 机器人学[M]. 北京:清华大学出版社,2000.
- [2] 张茂川,蔚伟,刘丽丽. 仿人机器人理论研究综述[J]. 机械设计与制造,2010(4):166-168.
- [3] 阮晓钢,仇忠臣,关佳亮. 双足行走机器人发展现状及展望[J]. 机械工程师,2007(2):17-19.
- [4] 谢涛,徐建峰,张永学,等. 仿人机器人的研究历史、现状及展望[J]. 机器人,2002,24(4):862-870.

(上接第65页)

法在实际的应用确实有效。表3是具体的对比统计数据。

表3 统计对比数据

| 营销主题     | 原算法聚类数量 | 改进算法聚类数量 | 原算法推广成功率 | 改进算法推广成功率 |
|----------|---------|----------|----------|-----------|
| 集团短号推广   | 10      | 13       | 23%      | 28%       |
| 套餐细分推广   | 10      | 16       | 30%      | 56%       |
| 竞争对手客户推广 | 10      | 13       | 33%      | 48%       |
| 客户流失特征分析 | 10      | 12       | 39%      | 57%       |

#### 4 结束语

文中给出K-means算法在移动通信行为特征分析客户聚类实例中的应用模式,改进了变量的量纲、维度、聚类数、初始聚点等关键影响因子的计算方法,通过具体项目的实践得到了预期的效果。但影响因子的计算方法不具备普遍适用性,研究具备普遍适用性的影响因子计算方法是下一步工作的重点。

#### 参考文献:

- [1] 范英,张忠能,凌君逸. 聚类方法在通信行业客户细分中的应用[J]. 计算机工程,2004,30(12):34-38.
- [2] 王军. 移动通信客户细分研究[D]. 北京:北京科技大学,2006:25-34.
- [3] 刘蓉,陈晓红. 基于数据挖掘的移动通信客户消费行为

- [5] 梶田秀司. 仿人机器人[M]. 北京:清华大学出版社,2007.
- [6] 张荣松,包家汉. 基于改进遗传算法的机器人路径规划[J]. 计算机技术与发展,2009,19(7):20-23.
- [7] Niku S B. Introduction to Robotics Analysis, Systems, Applications[M]. [s.l.]: Publishing House of Electronics Industry,2006.
- [8] Kajita S,Morisawa M,Harada k,et al. Biped Walking Pattern Generator allowing Auxiliary ZMP Control[C]//Proc. of 2006 IEEE/RSJ Intelligent Robots and Systems. [s.l.]:[s.n.],2006:2993-2999.
- [9] Harada K,Kajita S, Kanehiro F. et al. Real-Time Planning of Humanoid Robot's Gait for Force Controlled Manipulation [C]// Proc. of ICRA. [s.l.]:[s.n.],2004:616-622.
- [10] Tan Min, Xu De, Hou Zengguang. Advanced Robot Control [M]. Beijing:Higher Education Press,2007.
- [11] Kagami S,Kanehiro F,Tamiya Y,et al. AutoBalancer: An On-line Dynamic Balance Compensation Scheme for Humanoid Robots [C]//Proc. Int. Workshop Alg. Found. Robot. [s.l.]:[s.n.],2000.
- [12] Kajita S,Kanehiro F,Kaneko K,et al. Biped Walking Pattern Generation by Using Preview Control of Zero-Moment Point [C]// Proc. of 2003 Robotics and Automation. [s.l.]:[s.n.],2003:1620-1626.

分析[J]. 计算机应用与软件,2006,23(2):60-62.

- [4] 汪中,刘贵全,陈恩红. 一种优化初始中心点的K-means算法[J]. 模式识别与人工智能,2009,22(2):153-157.
- [5] 周世兵,徐振源,唐旭清. K-means算法最佳聚类数确定方法[J]. 计算机应用,2010,30(8):145-148.
- [6] 晋幼丽,周明全,王学松. SVM和K-means结合的文本分类方法研究[J]. 计算机技术与发展,2009,19(11):35-37.
- [7] Soper E,Suaanne F. The evolution of segmentation methods in services;where next[J]. Journal of Financial Services Marketing,2002,21(8):68-69.
- [8] 周卫星,廖欢. 基于K均值聚类和概率松弛法的图像区域分割[J]. 计算机技术与发展,2010,20(2):68-70.
- [9] Pawan L,Mofrh H. Temporal interval set analysis of clusters of supermarket customers;conventional approcah[J]. Information Sciences 2005,33(3):215-240.
- [10] Georgios P,Dimitrios P. The K-means range algorithm for personalized data clustering in e-commerce [J]. European Journal of Operation Research,2007,20(2):177-183.
- [11] 赵敏,倪志伟,刘斌. K-means与朴素贝叶斯在商务智能中的应用[J]. 计算机技术与发展,2010,20(4):179-182.
- [12] 刘英姿,吴昊客. 客户细分研究方法综述[J]. 管理工程学报,2006,20(1):53-57.