第24卷第4期 管理科学 ISSN 1672-0334 Vol. 24 No. 4 54-61
2011年8月 Journal of Management Science August, 2011

移动商务中面向客户细分的KSP混合聚类算法

邓晓懿[1],金 淳[1],樋口良之[2],韩庆平[3]
1 大连理工大学 系统工程研究所,辽宁 大连 116024
2 福岛大学 理工学部共生系统工程系,福岛 9601296
3 佛罗里达州立大西洋大学 信息技术及运作管理系,佛罗里达 博卡拉顿 33431

摘要:数据挖掘技术中的聚类算法是解决客户细分问题的重要算法之一。为解决传统聚类算法在客户细分问题中分类精度较低、收敛速度较慢的问题,着重对比分析传统聚类算法中K-means、自组织映射网络和粒子群3种算法的不足,提出融合3种算法优点的混合型聚类算法,该算法利用K-means和自组织映射网络对初始聚类中心进行优化,结合粒子群优化和K-means优化聚类迭代过程,并在迭代优化过程中设计避免算法因早熟而停滞的机制。针对移动电子商务环境下的餐饮业客户细分问题,建立移动餐饮业客户细分模型,并利用混合型聚类算法、K-means、层级自组织映射网络和基于粒子群的K-means等4种算法对实际案例进行对比分析。研究结果表明,混合型聚类算法的聚类精度分别比其他3种算法高,同时还具有最快的收敛性能,更适用于客户细分问题。

关键词:客户细分;K-means;自组织映射;粒子群优化;混合聚类
中图分类号:F713.5 **文献标识码**:A **文章编号**:1672-0334(2011)04-0054-08

1 引言

随着移动电子商务的发展,出现了使用移动终端进行消费的移动客户,且客户种类较多。由于情境和客户个性化特征不同[1],移动客户与普通客户间差异较大,如何合理地对客户进行细分已成为企业亟待解决的关键问题[2]。当前,使用聚类分析对客户进行细分管理是国内外研究热点[3],其中较具代表性的有K-means[4]、自组织映射(self-organizing feature map, SOM)网络和粒子群优化(particle swarm optimization, PSO)[5],它们在客户细分问题中得到广泛应用[6],如期货市场[7]、电力市场[8]和多媒体产品市场的客户细分[9]。

通常,客户细分问题对算法的求解精度要求较高。为更好解决客户细分问题,大量研究从传统聚类算法的初始化过程入手对算法进行改进,如Chiu等[10]、Niknam等[11]、Redmond等[12]和Laszlo等[13]的研究;部分研究对传统聚类算法求解过程进行优化,以期提高算法的聚类精度,如GKA[8]、HSOS[9]和WK-means[14]等算法。但大多数改进算法仅限于对局部过程和局部问题进行改进,使算法仍然存在易陷于局部最优[15]、过早收敛、聚类精度和客户细分效果提升幅度不大等问题。因此,提高客户细分算法精度是当前亟待解决的问题。针对移动环境下的客户细分研究较少[7],对移动环境下客户进行有效分类的衡量指标模型或细分方法相对匮乏。

本研究提出融合K-means、自组织映射网络和粒子群优化3种算法优点的混合型聚类算法,以有效提高算法的聚类精度,结合移动商务环境下的餐饮客户细分模型分析并验证算法的有效性。

2 相关研究评述

2.1 K-means

K-means算法是由MacQueen[4]提出的一种基于划分的聚类分析方法,算法流程描述如下。

收稿日期:2010-11-16 **修返日期**:2011-04-18
基金项目:国家自然科学基金(70890080,70890083)
作者简介:邓晓懿(1982-),男,辽宁大连人,大连理工大学系统工程研究所博士研究生,研究方向:电子商务与数据挖掘、信息管理和系统优化等。E-mail:pltdeng@dlut.edu.cn

设含有 m 个数据的 n 维样本点集 $D=\{x_i \mid i=1,2,\cdots,m\}$,点 $x_i=(x_{i1},x_{i2},\cdots,x_{in})^T$,$K$ 为聚类(簇)数目,C 为聚类中心集合,$C=\{c_j \mid j=1,2,\cdots,K\}$,$c_j=(c_{j1},c_{j2},\cdots,c_{jn})^T$,$d(x_i,c_j)$ 为点 x_i 到聚类中心 c_j 的欧式距离,有

$$d(x_i,c_j)=\sqrt{\|x_{i1}-c_{j1}\|^2+\|x_{i2}-c_{j2}\|^2+\cdots+\|x_{in}-c_{jn}\|^2} \quad (1)$$

步骤 1　从 D 中随机选择 K 个点作为初始聚类中心 C;

步骤 2　计算并比较每个点 x_i 到聚类中心 c_j 的距离 $d(x_i,c_j)$,将 x_i 分配到与其距离最小的簇中;

步骤 3　重新确定新的聚类中心 C;

步骤 4　重复执行步骤 2 和步骤 3,直至聚类中心 C 不再发生变化。

K-means 算法的优点是效率高、复杂度低以及易于实现,但也存在以下两点缺陷。

(1)对初始聚类中心敏感,对于不同的初始中心,可能导致不同的聚类结果,直接影响聚类效果;同时,要求事先设定 K 值,而在实际中 K 值直接决定了聚类的数目,难以确定。

(2)基于梯度下降法,可能会过早地收敛于局部最优解,即全局搜索能力较差。

针对 K-means 对初始聚类中心敏感的缺点,Redmond 等[12]利用 *kd* -trees 对算法进行改进,克服了需要初始 K 值的不足;Laszlo 等[13]提出基于遗传算法的 K-means 聚类,改进 K 值初始化过程。为克服 K-means 算法易过早收敛的不足,Huang 等[14]优化了 K-means 聚类过程中的变量选择方式,以提高算法的全局解搜索能力;Carpaneto 等[8]利用 K-means 算子代替 GA 中的交叉算子,弥补算法容易收敛于局部最优解的缺点。

以上研究对 K-means 算法修改的局限性在于,或者改进算法的初始化过程,或者只优化算法对全局最优解的搜索能力,并没有同时从以上两方面进行综合改进,难免会造成聚类精度和客户细分效果提升幅度不大等问题。

2.2 SOM

SOM 网络是一种被广泛应用于聚类、具有自组织功能的无监督学习神经网络[6],其聚类方法是为每个输入神经元寻找对应最佳的输出神经元,基本描述如下。

设 SOM 网络含有 n 个输入神经元和 N 个输出神经元,输入神经元集合为 $(x_1,x_2,\cdots,x_n)$,输出神经元集合为 $(u_1,u_2,\cdots,u_N)$,W 为神经元间的连接权值集合,$W=\{W_{ij} \mid i=1,2,\cdots,n,\ j=1,2,\cdots,N\}$。SOM 聚类就是寻找与输入神经元 x_i 距离最近的输出神经元 u_j 以及相应的连接权值 W_{best},即满足

$$\|x_i-W_{best}\| \leqslant \|x_i-W_{ij}\|,\ \forall i \in n,\ \forall j \in N \quad (2)$$

SOM 网络具有自稳定性、无须定义评价函数、能够识别向量空间中最有意义特征等优点,对于一般客户细分问题,SOM 的聚类效果较好,但 SOM 需要预先构建网络结构,而且训练样本的输入次序会影响聚类结果[6]。对此不足,Hung 等[9]提出一种层次 SOM 的客户细分方法,但该方法的不足在于使用随机方式初始化 SOM 权值向量,造成 SOM 网络构建和训练时间较长[10]。

2.3 PSO

PSO 是一种通过群体中粒子间的合作和竞争产生的优化算法[5],算法表述如下。

设 n 维空间中数目为 m 的粒子群 $X=\{x_i \mid i=1,2,\cdots,m\}$,每个粒子 x_i 有速度 $V_i=(v_{i1},v_{i2},\cdots,v_{in})$ 和位置 $P_i=(p_{i1},p_{i2},\cdots,p_{in})$ 两种属性,v_{in} 和 p_{in} 分别为 x_i 在第 n 维空间中的速度和位置。通过(3)式和(4)式不断调整自身速度 V_i 和位置 P_i 来搜索新解,在搜索过程中每个粒子都能记录下自身搜索到最优解 P_i^{best} 和整个粒子群已搜索到的最优解 G_{best}。

$$V_i^{t+1}=\omega V_i^t+c_1R_1(P_i^{best}-P_i)+c_2R_2(G_{best}-P_i) \quad (3)$$

$$P_i^{t+1}=P_i+V_i^{t+1},\ i=1,2,\cdots,m \quad (4)$$

其中,ω 为惯性权重,c_1 和 c_2 为加速度常数,R_1 和 R_2 为服从 $U(0,1)$ 分布的随机数,t 为当前迭代次数。

PSO 比 GA 具有更快的收敛速度,但依然具有一定的早熟倾向,造成其求解全局最优解的精确度较低。为使 PSO 更好地适应客户细分问题,Chiu 等[10]提出基于 PSO 和 K-means 的 PSOK 算法,Niknam 等[11]和 Firouzi 等[15]分别提出基于 PSO、ACO 和 K-means 以及基于 PSO、SA 和 K-means 的组合聚类算法,提高了原算法的求解精度。但这些研究只是利用 PSO 与 ACO 或者 PSO 与 SA 的组合对原算法的初始化过程进行优化,并没有解决单个算法易收敛于局部最优而造成全局解精确度降低的问题。

针对上述客户细分算法改进的不足,本研究在上述工作的基础上,提出混合聚类算法(hybrid clustering algorithm based on K-means, SOM and PSO, KSP),从初始化和求解过程两方面对客户细分算法进行改善。

3 KSP 混合聚类算法

3.1 KSP 算法设计思路

3.1.1 改进算法初始化过程

首先,针对 K-means 算法对初始聚类中心敏感的缺点,使用 SOM 网络对客户数据进行预处理,将得到的聚类中心作为下一步聚类的初始聚类中心。同时,为克服 SOM 网络构建和训练时间较长的缺点,利用 K-means 算法对 SOM 网络的训练数据进行预聚类,利用预聚类数据初始化 SOM 网络权值,以减少 SOM 网络训练时间,加快 SOM 网络训练过程。具体改进措施如下。

(1)从待处理数据集 V 中选取部分样本 V' 作为训练数据集,使用 K-means 将 V' 划分为 K 个簇,得到聚类中心集 C;

(2)使用 C 初始化具有 n 个输入神经元和 N 个输出神经元的 SOM 网络的神经元权值集合 W,$N=n\cdot n$,详细描述如下。

①初始化处于 SOM 网络顶点位置的神经元权值。先从 C 中寻找间距最大的两个聚类中心点 c_i 和

c_j, $i,j=1,2,\cdots,K$ 且有 $i\neq j$,将它们分别作为对角线上两个顶点神经元的初始化权值;然后从剩余的聚类中心集合 $C-\{c_i,c_j\}$ 中寻找与 c_i 和 c_j 间距最大的聚类中心点 c_k, $k\in K-\{i,j\}$,将 c_k 作为初始化权值赋予处于副对角线上的任意一个顶点神经元;最后从 $C-\{c_i,c_j,c_k\}$ 中寻找与 c_i、c_j 和 c_k 间距最大的聚类中心点 c_l, $l\in K-\{i,j,k\}$,将它赋予处于最后一个顶点神经元。

②初始化 SOM 网络外层 4 边的神经元权值。将外层 4 边表示为上下左右 4 部分,以左侧边为例介绍初始化方法。首先,通过(1)式依次计算剩余 $K-4$ 个聚类中心分别与左侧边最上方两个端点间的距离之和,从中选取数值最小的 $n-2$ 个聚类中心;然后,将这些聚类中心作为初始化权值,按照与左上顶点距离递增的顺序从左至右依次赋予各神经元,其余各边的处理方法相同。

③由外向内逐层初始化剩余神经元权值。方法与①和②初始化最外层神经元权值的步骤类似,但需要注意一点,即找到初始化每层的左上角和右下角的两个点后,用两点中与网络最外层左上角距离最近的点来初始化该层左上角的神经元权值,其他的与上面的步骤完全相同。

由于 SOM 网络的权值分布与样本数据分布直接相关,当样本数据离散程度较高且样本分布与权值分布保持一致时,SOM 网络收敛速度最快[9]。因此,本研究提出的上述步骤先对预处理后的样本数据进行离散化处理,然后直接将该数据作为 SOM 网络的初始权值,以保证初始权值分布与样本分布的一致性,最终可提高 SOM 网络的收敛速度。

(3)使用样本数据 V' 训练 SOM 网络,用训练好的 SOM 网络对待处理数据 V 进行分类,将 V 划分成 K 个簇,得到聚类中心集合 C', $C'=\{C'\mid j=1,2,\cdots,K\}$,并将 C' 作为算法下一步过程的初始化数据。

3.1.2 优化算法求解过程

PSO 在产生下一代群体时带有一定随机性,保持了粒子的多样性,同时也扩大了 PSO 的全局最优搜索范围,使其全局解精确度较低。针对这一点,本研究将 K-means 和 PSO 两者相结合,在 PSO 更新粒子个体时,利用 K-means 的局部最优解搜索力较强的特性,使用 K-means 对 PSO 历代产生的新粒子个体进行聚类优化,以提高算法全局最优解的搜索能力。同时,还在迭代优化过程中设计了避免算法因早熟而停滞的机制,可有效使算法跳出局部最优,减少无效迭代次数,从而提高算法的收敛速度。优化措施如下。

(1)修正惯性权重 ω。ω 值直接决定 PSO 的搜索范围[16],如果 ω 大,则适用于全局搜索,收敛速度快,但求解精度低;如果 ω 小,则有利于局部搜索,求解精度高,但收敛速度慢。由于 PSO 算法的搜索特点为前期速度快、后期速度慢,且为非线性下降,为使算法具有较强的全局搜索能力,同时在搜索前期能保持较高的搜索效率,在后期又能保持较高的搜索精度,本研究对 ω 进行修正,即

$$\omega=\omega_{\max}-(\omega_{\max}-\omega_{\min})\left(\frac{t}{t_{\max}}\right)^r \tag{5}$$

其中,$\omega_{\max}$ 为惯性权重的最大值,$\omega_{\min}$ 为惯性权重的最小值,t 为当前迭代次数,$t_{\max}$ 为最大迭代次数,r 为大于 2 的自然数,视具体情况而定。

在对 ω 进行修正后,PSO 的粒子速度更新(3)式变更为(6)式,即

$$V_i^{t+1}=\left[\omega_{\max}-(\omega_{\max}-\omega_{\min})\left(\frac{t}{t_{\max}}\right)^r\right]V_i^t+c_1R_1(P_i^{best}-P_i)+c_2R_2(G_{best}-P_i) \tag{6}$$

(2) 使用 K-means 优化新群体,可分为以下两步。

①根据最近原则将新一代粒子添加到距离其最近聚类中心 C' 所属的簇中,然后使用 K-means 对形成的粒子簇进行聚类,得到新的聚类中心集合 C'', $C''=\{C''_j\mid j=1,2,\cdots,K\}$。

② 根据(7)式计算粒子群中粒子 x_i 在其当前位置的适应度 $F(P_i)$,分别比较 $F(P_i)$ 与粒子经历过的最好位置 P_i^{best} 的适应度 $F(P_i^{best})$ 和粒子群经历过的最好位置 G_{best} 的适应度 $F(G_{best})$,之后用适应度较大的位置更新粒子的速度并调整其位置。最后,更新整个粒子群经过的最好位置 G_{best}。

$$F(P_i)=\frac{\lambda}{\sum_{j=1}^{K}\sum_{P_i\in C_j}\|P_i-C_j''\|}\quad i=1,2,\cdots,m,\ j=1,2,\cdots,K \tag{7}$$

其中,C_j'' 为 C'' 中第 j 个簇的聚类中心,P_i 属于 C_j'', $\sum_{j=1}^{K}\sum_{P_i\in C_j}\|P_i-C_j''\|$ 为聚类内离散度之和,λ 为常数,m 为粒子数。显然,粒子的适应度与离散度之和成反比,即聚类内离散度之和越小,粒子的适应度越大。

(3) 判断并阻止算法陷入局部最优。由(3)式和(4)式可知,当前最优解 P_i^{best} 和全局最优解 G_{best} 同时影响粒子运动方向。假设 P_i^{best} 和 G_{best} 同为局部最优值,此时,粒子会受 P_i^{best} 和 G_{best} 的影响,进入重复、相同的搜索路径,即陷入局部最优。虽然(6)式可以根据 PSO 的搜索特点改变粒子在不同阶段的搜索步长,但并不能改变粒子的运动方向,无法从根本上使算法跳出局部最优。

在算法迭代过程中,如果当前全局最优解 G_{best} 在连续若干次迭代后依然不变时,表明算法即将或已经陷于局部最优解,判断算法是否出现早熟的阈值 T_p 则可以根据求解问题规模或数值实验测定。此时,为使粒子跳出局部最优值,对其当前最优解 P_i^{best} 的任意一维数据进行小幅修改,强迫改变粒子的下一步运行方向,即

$$P_i^{best\,\prime}=(\boldsymbol{E}+\boldsymbol{\varepsilon})\,P_i^{best},\ i=1,2,\cdots,m \tag{8}$$

其中,$\boldsymbol{E}$ 为 n 维单位向量;$\boldsymbol{\varepsilon}$ 为方向修正权重,为 n 维向量,且有 $\boldsymbol{\varepsilon}\sim U(0,1)$。一旦发现算法出现早熟,通过该方法对算法当前的全局最优解进行随机微调,可有效地减少无效迭代次数,提高算法收敛速度,从而

提高聚类精度。

3.2 KSP 算法流程及复杂度

3.2.1 算法流程

Begin

输入 V、V', SOM 网络输入/输出神经元数为 n、N,神经元权值 W,粒子群 P,聚类数 K,阈值 T_p,迭代记数 $T=0$

使用 K-means 对 V' 进行初始聚类,得到预聚类中心集 C

根据3.1.1 中算法初始化步骤,使用 C 初始化 SOM 网络的神经元权值集合 W

使用 V' 对 SOM 网络进行训练

使用训练完毕的 SOM 对 V 进行聚类,得到聚类中心集合 C'

使用集合 C' 初始化粒子群 P,将 C' 作为各粒子的初始位置集合

do{

根据(7)式计算粒子 P_i 的适应度 $F(P_i)$

if $F(P_i) > F(P_i^{best})$

更新 P_i^{best}

endif

if $F(P_i) > F(G_{best})$

更新 G_{best}

endif

根据(4)式和(6)式计算粒子 P_i 的速度和位置

使用 K-means 对新生成的新一代粒子群进行聚类,得到新的聚类中心 C''

计算粒子 P_i 的适应度 $F(P_i)$

if $F(P_i) > F(P_i^{best})$

更新 P_i^{best}

endif

if $F(P_i) = F(G_{best})$

$T = T+1$

endif

if $F(P_i) > F(G_{best})$

更新 G_{best}

endif

if $T = T_P$

根据(8)式更新 P_i^{best}, $T=0$

endif

重新计算粒子 P_i 的速度和位置

记录 P_i^{best} 和 G_{best}

}

while (P_i^{best} 和 G_{best} 不稳定)

输出全局最优解及历代最优解

End

3.2.2 算法复杂度

通常状况下,K-means 算法时间复杂度为 $O(n)$[4];SOM 网络的时间复杂度为 $O(T_1nN)$, T_1 为 SOM 实际迭代次数, n 和 N 分别为输入/输出神经元个数[9];PSO 的时间复杂度为 $O(T_2PD)$, T_2 为 PSO 实际迭代次数, P 为粒子个数, D 为粒子的维度[11]。三者的时间复杂度对比为 $O(n) < O(T_1nN) \approx O(T_2PD)$。本研究提出的 KSP 算法首先运行 K-means 和 SOM,之后在 PSO 的迭代过程中引入 K-means 算法,因此 KSP 的时间复杂度为 $O(\max(T_1nN,\ nT_2PD))$。与前 3 种算法相比,KSP 与三者中时间复杂度最高者相同。

4 基于 KSP 的客户细分模型

客户细分通常采用 RFM 模型[16](recency/frequency/monetary model, RFM)和 LTV 模型[17](lifetime value model, LTV),在餐饮业客户细分问题中存在相当数量使用移动终端进行消费的移动客户。根据 RFM 模型可知,该部分客户的近度(R)趋于零、频度(F)较高,且不同客户间的区分度趋于零,因而在移动环境下单纯使用 RFM 模型分析餐饮业客户的意义较小,需结合其他指标对客户细分。

此时,从 RFM 模型中的频度(F)、LTV 模型中的客户消费额和利润三方面可以较为直观地区分普通客户和移动客户。因此,本研究结合 RFM 和 LTV 对移动电子商务环境下的餐饮客户细分问题进行分析,提出移动环境下餐饮客户价值评价指标模型。

4.1 建立客户评价指标体系

4.1.1 客户评价指标的选择

本研究结合 RFM 和 LTV 两种模型,定义消费指标 S、利润指标 P 和客户指标 C 对客户价值进行评价,如表 1 所示。消费指标和利润指标描述客户的现实价值,客户指标反映移动环境下的客户消费模式以及客户潜在价值。由于部分客户指标涉及到客户隐私,获取可行性较低,而且也无法将指标量化,所以按照定性指标评分方法进行处理。

4.1.2 客户评价指标的确定

由表 1 可以看出,消费指标 S、利润指标 P、客户指标 C 确定了客户价值,同时也是进行客户细分的依据,它们分别由其下层相应的各项三级指标决定,无法直接获取。因此,可利用层次分析法(analytic hierarchy process, AHP)确定客户在消费指标、利润指标、客户指标上的取值。大致步骤如下。

(1)确定三级指标对二级指标的权重。根据 AHP 法,邀请专家组成评价小组,根据各因素相对上层指标的相对重要程度建立判断矩阵 W,计算 W 的特征值 $\tilde{W}$ 作为各三级指标对其二级指标的权重[18]。

(2)确定三级指标。在确定三级指标时,对其中的定量指标和定性指标分别进行评价。对于定量指标,可根据极差法对数据进行标准化处理,即

$$R_{ij} = \frac{P_{ij} - P_{\min}}{P_{\max} - P_{\min}} \tag{9}$$

其中, R_{ij} 为客户 j 在指标 i 上的表现值, P_{ij} 为待评价客户 j 在指标 i 上的得分值, $P_{\max}$ 为待评价客户在指标 i 上的最大表现值, $P_{\min}$ 为待评价客户在指标 i 上的最小表现值。

表 1 移动环境下的餐饮客户价值评价指标体系

Table 1 Evaluation Index System for Customer Value of Catering Enterprise in Mobile Environment

一级指标	二级指标	三级指标	指标描述	指标类型
餐饮客户价值评价指标	消费指标 S	客户消费额	一定时间内客户的消费额	定量
		客户消费频率	一定时间内客户的消费频率	定量
	利润指标 P	客户利润	一定时间内获取的客户利润	定量
		客户利润率	一定时间内获取的客户利润与客户消费额的比率	定量
	客户指标 C	移动消费份额	一定时间内客户使用移动终端的消费额与总消费额的比例	定量
		移动消费利润率	一定时间内客户使用移动终端消费利润率与总利润的比率	定量
		移动消费频率比	一定时间内客户使用移动终端消费次数与总消费次数的比例	定量
		客户年龄	描述客户的年龄阶段	定性

对于定性指标,采用专家打分法确定各指标的取值以及打分等级,然后将专家对客户的评价按相应的分值加权平均后作为客户在此指标上的分值。

(3)确定二级指标。将各三级指标的得分值加权平均后可得到各二级指标的分值,即

$$R_i^2 = \tilde{W}_i(R_{i1}^3, R_{i2}^3, \cdots, R_{ij}^3) \tag{10}$$

其中,$\tilde{W}_i$ 为权重,R_i^2 为第 i 个二级指标,R_{ij}^3 为属于 R_i^2 的第 j 个三级指标。

本研究邀请某大型餐饮企业的两位经理和 5 位业内相关人士,结合近千份有效客户问卷对比分析各三级指标对各二级指标 S、P 和 C 的权重,最终得到的权重值为

$$\tilde{W} = \begin{cases} \tilde{W}_S = (0.90, 0.10) \\ \tilde{W}_P = (0.90, 0.10) \\ \tilde{W}_C = (0.40, 0.11, 0.40, 0.09) \end{cases}$$

4.2 应用 KSP 对客户细分

在得到客户评价指标后,应用 KSP 算法对客户进行分类,基本流程如下。

(1)使用 KSP 算法对客户的 S、P、C 指标进行聚类,得到 K 类客户群体。

(2)比较每个客户簇的 S、P、C 的平均值 $\overline{S}_i$、$\overline{P}_i$、$\overline{C}_i$ 和所有客户的 S、P、C 总均值 $\overline{S}$、$\overline{P}$、$\overline{C}$,每次对比分为大于等于和小于两种情况,通过对比可得到每个客户簇 S、P、C 指标的变动情况。如果单个簇的均值大于等于总均值,便给该值一个“+”标记,反之则标记“-”。

(3)根据每个客户簇的 S、P、C 指标变动情况分析该类客户的性质,定义客户类型。

(4)使用 KSP 算法对客户的 S、P、C 指标进行聚类,得到 K 类客户群体。

5 案例分析和算法对比

5.1 案例分析

目前,中国的移动电子商务正处于发展阶段,移动环境下提供相关服务的企业相对较少。鉴于此,本研究选取日本某中型连锁餐厅作为案例企业。随着移动电子商务的普及,该餐厅提供多种订餐方式,如电话、Internet、Email 和移动网络,客户类型较多,客户管理难度较大。本研究从近年在餐厅就餐的客户中选取6 000 位作为数据样本,应用基于 KSP 算法的客户细分模型对客户数据进行分析,得到 9 个客户簇,如表 2 所示。

对比表 2 中 S、P、C 的变化情况,客户类型可归纳为 5 类,即簇 1、簇 3、簇 6、簇 9 为同一类型,设为类型 A;簇 2 为一种类型,设为类型 B;簇 4 设为类型 C;簇 5 设为类型 D;簇 7 和簇 8 设为类型 E。不同客户类型的识别说明如表 3 所示,可以看出,移动客户(52.65%)略多于普通客户(47.35%),交易额和利润额较高的忠实移动客户占全体客户的27.35%,而大部分普通客户属于一般客户和不确定客户。

5.2 算法对比

通过本研究的 KSP 算法与 K-means、基于 SOM 的 HSOS[10]、基于 PSO 和 K-means 的 PSOK[11] 3 种既有算法对案例数据的聚类分析结果的比较,验证 KSP 的聚类精度和效率。根据已有研究[15],选取聚类间的离散度之和,即所有数据点到其最近聚类中心的欧氏距离之和作为适应度函数;聚类精度(*Accuracy*)则使用聚类内离散度(*Intra_Clustter-D*)和聚类间离散度(*Inter_Clustter-D*)的比值来描述,具体为

$$Perf(X) = \sum_{j=1}^{k} \sum_{X_i \in C_j} \| X_i - C_j \| \tag{11}$$

表 2　KSP 聚类结果

Table 2　Results of KSP Clustering

簇编号	客户数量	消费指标 (S)	利润指标(P)	客户指标(C)	指标比较
1	311	0.476	0.325	0.814	$S+P+C+$
2	602	0.201	0.147	0.443	$S+P-C-$
3	184	0.499	0.559	0.795	$S+P+C+$
4	1 231	0.131	0.098	0.465	$S-P-C+$
5	287	0.175	0.263	0.498	$S-P+C+$
6	173	0.836	0.728	0.986	$S+P+C+$
7	1 312	0.029	0.012	0.244	$S-P-C-$
8	927	0.078	0.061	0.264	$S-P-C-$
9	973	0.289	0.239	0.629	$S+P+C+$
	总均值	0.183	0.150	0.453	

表 3　客户类型识别结果

Table 3　Results of Customer Classification

序号	类型	簇编号	数量	指标比较	客户类型说明
1	A	1,3,6,9	1 641	$S+P+C+$	移动客户,交易额和利润高,可视为餐厅的忠实移动客户
2	B	2	602	$S+P-C-$	普通客户,交易额高,利润较低,可能对价格比较敏感,可视为一般客户
3	C	4	1 231	$S-P-C+$	移动客户,交易额和利润较低,可视为一般移动客户
4	D	5	287	$S-P+C+$	移动客户,交易额低,利润较高,可视为待发展的潜在优质客户
5	E	7,8	2 239	$S-P-C-$	普通客户,交易额和利润低,为无价值客户或新客户,可视为待观察客户

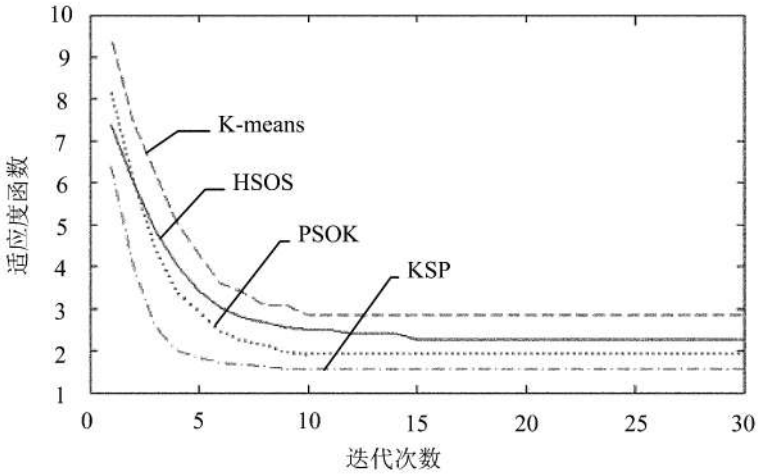

图 1　算法收敛速度对比

Figure 1　Comparison of Algorithms′ Convergence Rate

$$Intra\_Cluster\text{-}D=\sum_{i=1}^{n}\sum_{j=i+1}^{n}\|x_i-x_j\| \tag{12}$$

$$Inter\_Cluster\text{-}D=\sum_{i=1}^{k}\sum_{j=i+1}^{k}\|C_i-C_j\| \tag{13}$$

$$Accuracy=\left(1-\frac{Intra\_Cluster\text{-}D}{Inter\_Cluster\text{-}D}\right)\times 100\%$$

$$=\left(1-\frac{\sum_{i=1}^{n}\sum_{j=i+1}^{n}\|x_i-x_j\|}{\sum_{i=1}^{k}\sum_{j=i+1}^{k}\|C_i-C_j\|}\right)\times 100\% \tag{14}$$

图 1 给出 4 种算法适应度$Perf(X)$的收敛曲线,横坐标为算法的迭代次数,纵坐标为算法的适应度

表4 算法运行时间和聚类精度比较

Table 4 Comparison of Running Time and Clustering Accuracy

算法名称	运算时间(ms)	*Intra_Clustter-D*	*Inter_Clustter-D*	*Accuracy*(%)
K-means	15	0.247	0.569	56.591
HSOS	8 750	0.212	0.597	64.489
PSOK	4 738	0.216	0.632	65.822
KSP	5 129	0.195	0.664	70.633

值。在同等条件下,KSP 的收敛速度最快,PSOK 次之,K-means 的收敛速度较 PSOK 稍慢,而 HSOS 的收敛速度最慢。

4 种算法的聚类精度和运行时间对比见表 4。KSP 的聚类精度为 70.633%,比 K-means、HSOS 和 PSOK 这 3 种算法分别高14.042%、6.144%和4.811%,为各种算法中最高;在运行时间方面,KSP 的运行时间介于 PSOK 和 HSOS 之间,说明 KSP 的时间复杂度与 K-means、HSOS 和 PSOK 三者中最高者相同,K-means 的运行时间最短,HSOS 的运行时间最长。在实际应用中,客户细分问题往往要求算法具有较高求解精度,对于运行时间的要求较低。因此,从求解精度方面看,KSP 算法优于其他 3 种现有算法。

6 结论

在移动商务中,由于情境和客户个性化特征不同,造成不同客户间的差异较大,导致传统聚类算法在客户细分问题中的精度较低,而且移动环境下的客户细分模型相对较少。针对移动商务中的客户细分问题,本研究以 K-mean、SOM 和 PSO 算法为基础,提出一种融合 3 种算法优点的混合型聚类算法,即 KSP 算法,并建立一个移动商务环境下的餐饮客户价值评价模型,结合 KSP 算法对移动餐饮客户进行细分。

通过实际案例的验证结果发现,与其他 3 种聚类算法相比,KSP 算法不仅在聚类精度方面优于 K-means、HSOS 和 PSOK 算法4.811% ~14.042%,还具有最快的收敛速度,改进了传统聚类算法对初始中心敏感、易陷于局部最优解的不足,更适用于客户细分问题。未来的研究将从算法的执行效率方面入手,减少算法的运行时间,使之更适用于实时、动态的客户细分问题。

参考文献:

[1] 胡理增,薛恒新,于信阳. 以客户终身价值为准则的客户重要程度识别系统[J]. 系统工程理论与实践,2005,25(11):79-85.
Hu Lizeng, Xue Hengxin, Yu Xinyang. The customer importance identification system based on customer lifetime value [J]. Systems Engineering-Theory & Practice, 2005, 25(11): 79-85. (in Chinese)

[2] 夏维力,王青松. 基于客户价值的客户细分及保持策略研究[J]. 管理科学,2006,19(4):35-38.
Xia Weili, Wang Qingsong. Customer segmentation and retention strategy based on customer value [J]. Journal of Management Science, 2006, 19(4): 35-38. (in Chinese)

[3] Böttcher M, Spott M, Nauck D, Kruse R. Mining changing customer segments in dynamic markets [J]. Expert Systems with Applications, 2009, 36(1): 155-164.

[4] MacQueen J. Some methods for classification and analysis of multivariate observations [C] // Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probabilit. Berkeley: University of California Press, 1967: 281-297.

[5] 刘晓峰,陈通,张连营. 基于微粒群算法的最佳证券投资组合研究[J]. 系统管理学报,2008,17(2):221-224,234.
Liu Xiaofeng, Chen Tong, Zhang Lianying. Study on the portfolio problem based on particle swarm optimization [J]. Journal of Systems & Management, 2008, 17(2): 221-224, 234. (in Chinese)

[6] Ngai E W T, Xiu L, Chau D C K. Application of data mining techniques in customer relationship management: A literature review and classification [J]. Expert Systems with Applications, 2009, 36(2): 2592-2602.

[7] 陈智高,陈月英,常香云. 基于客户价值的期货业客户聚类细分方法[J]. 清华大学学报:自然科学版,2006,46(S1):1046-1051.
Chen Zhigao, Chen Yueying, Chang Xiangyun. Customer clustering segmentation method for futures industry based on customer value [J]. Journal of Tsinghua University: Science and Technology, 2006, 46(S1): 1046-1051. (in Chinese)

[8] Carpaneto E, Chicco G, Napoli R, Scutariu M. Electricity customer classification using frequency: Domain load pattern data [J]. International Journal of Electrical Power & Energy Systems, 2006, 28(1): 13-20.

[9] Hung C L, Tsai C F. Market segmentation based on hierarchical self-organizing map for markets of multimedia on demand[J]. Expert Systems with Applications, 2008, 34(1): 780-787.

[10] Chiu C Y, Chen Y F, Kuo I T, Ku H C. An intelligent market segmentation system using k-means and particle swarm optimization[J]. Expert Systems with Applications, 2009, 36(3): 4558-4565.

[11] Niknam T, Amiri B. An efficient hybrid approach based on PSO, ACO and k-means for cluster analysis[J]. Applied Soft Computing, 2010, 10(1): 183-197.

[12] Redmond S J, Heneghan C. A method for initialising the K-means clustering algorithm using kd-trees [J]. Pattern Recognition Letters, 2007, 28(8): 965-973.

[13] Laszlo M, Mukherjee S. A genetic algorithm that exchanges neighboring centers for k-means clustering [J]. Pattern Recognition Letters, 2007, 28(16): 2359-2366.

[14] Huang J Z, Ng M K, Rong H, Li Z. Automated variable weighting in k-means type clustering[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(5): 657-668.

[15] Firouzi B B, Sadeghi M S, Niknam T. A new hybrid algorithm based on PSO, SA, and K-means for cluster analysis[J]. International Journal of Innovative Computing, Information and Control, 2010, 6(7): 3177-3192.

[16] Chen Y L, Kuo M H, Wu S Y, Tang K. Discovering recency, frequency, and monetary (RFM) sequential patterns from customers' purchasing data[J]. Electronic Commerce Research and Applications, 2009, 8(5): 241-251.

[17] Chai C, Chan H. Intelligent value-based customer segmentation method for campaign management: A case study of automobile retailer[J]. Expert Systems with Applications, 2008, 34(4): 2754-2762.

[18] 马东辉,郭小东,苏经宇,周锡元,钱稼茹. 层次分析法逆序问题及其在土地利用适宜性评价中的应用[J]. 系统工程理论与实践, 2007, 27(6): 124-135, 165.
Ma Donghui, Guo Xiaodong, Su Jingyu, Zhou Xiyuan, Qian Jiaru. Inverted-order problem and application in suitability assessment of land use for AHP[J]. Systems Engineering-Theory & Practice, 2007, 27(6): 124-135, 165. (in Chinese)

KSP: A Hybrid Clustering Algorithm for Customer Segmentation in Mobile E-commerce

Deng Xiaoyi[1], Jin Chun[1], Higuchi Yoshiyuki[2], Han Jim[3]

1 Institute of Systems Engineering, Dalian University of Technology, Dalian 116024, China

2 Faculty of Symbiotic Systems Science, Fukushima University, Fukushima 9601296, Japan

3 Department of Information Technology and Operations Management, Florida Atlantic University, Boca Raton 33431, USA

Abstract: Clustering algorithms in data mining technology is an important kind of algorithms of soloving customer segmentation problems. To overcome the low accuracy and slow convergence of traditional clustering algorithms in customer segmentation, this paper analyzes deficiencies of traditional cluster algorithms, K-means, SOM and PSO. After that, an improved hybrid clustering algorithm named KSP is proposed, which integrates advantages of K-means, SOM and PSO. The initialization of KSP is optimized by K-means and SOM; the solving process is carried out by the combination of PSO and K-means with a mechanism of restraining premature stagnancy. Then, a customer segmentation model was established to analyze types of customers in catering industry under mobile electronic commerce environment. Also, an actual case was illustrated to verify the efficiency of the KSP algorithm. The results show that the KSP has the highest accuracy and convergence rate. Thus, it is more suitable for customer segmentation.

Keywords: customer segmentation; K-means; self-organizing map; particle swarm optimization; hybrid clustering

Received Date: November 16th, 2010　　**Accepted Date**: April 18th, 2011

Funded Project: Supported by the National Natural Science Foundation of China(70890080, 70890083)

Biography: Deng Xiaoyi, a Liaoning Dalian native(1982 -), is a Ph. D. candidate in the Institute of Systems Engineering at Dalian University of Technology. His research interests include e-commerce and data mining, information management, system optimization, etc.

E-mail: pltdeng@dlut.edu.cn

□