一种基于动态模糊聚类算法的客户细分方法

A Kind of Client Market Segmentation Based on the Dynamic Clustering Algorithm

吴春旭　吴 镝　蒋 宁　（中国科学技术大学管理学院　安徽合肥 230052）

摘　要： 本文提出了基于信息熵和K均值算法混合迭代模糊聚类的客户细分模型，解决了模糊聚类的原型初始化参数问题。将信息熵和K均值算法引入模糊聚类中进行分析，并结合联通客户的大样本数据进行实际分析，与传统方法相比，取得了较好的效果。

关键词： 客户细分 模糊聚类 信息熵 K均值算法 FMC

1　引言

客户细分是以消费者需求为出发点，根据消费者购买行为的差异性，把消费者总体划分为类似性购买群体的过程[1]。目前客户细分研究根据导向视角不同主要形成两大流派，一派是消费者导向细分，主要为理论界采用，Yoram Wind[2]认为客户细分重点应当是对消费者需求和行为特征进行分类，其中由于RFM[3]的细分变量容易量化，应用中效果比其他细分方法明显。另一派是产品导向市场细分，主要为营销决策者采用，根据不同营销决策目标，围绕某产品或品牌特定消费情境对消费者细分。我国企业长期以来缺乏科学的客户细分和目标客户定位，目前的细分标准不科学，不能准确分析和掌握客户需求、态度和偏好，且客户细分方法大多是单维的，对市场竞争日益激烈的企业来说并不适用，而对于传统统计分类的属性变量不易收集和更新，基于动态模糊聚类算法的客户细分方法可以有效地解决这些问题，反映客户消费行为特性。

2　动态模糊聚类的初始化

2.1　信息熵理论及其在初始类个数确定中的应用

1948 年 C. E. Shannon 将熵概念引入信息理论，用信息熵来量度信息量，其表达式定义如下[4]

$$H(x) = -\sum_{i=1}^{r} P(a_i)\log P(a_i) \qquad (2.1)$$

a_i：信源端各个不同符号

$H(x)$：信源总体信息量的统计平均值

模糊聚类起源于 1965 年 L. A. Zadeh 创立的模糊集[5]。模糊聚类的结果是得到样本对于每个类的隶属程度，即属于类的不确定程度，着重考察的是样本点，但是对于每个类及整体分类结果的不确定程度却没有度量。由于聚类分析需要事先确定类个数，类个数很少时，虽然对于整个数据集的划分会较为清晰，但对于每个类的特性认知会较为笼统。而当初始类个数较多时，数据集的划分会较为细碎，虽然每个类的特性认知会较为清晰，但总体特征模糊。

$$p_{ij} = \frac{d_{ij}}{\sum_{k=1}^{n} d_{kj}} \qquad (2.2)$$

d_{ij}：第 i 个样本点对于第 j 个类中心距离

P_{ij}：第 i 个样本点对于第 j 个类中心的偏离度

偏离度：设共有 n 个样本点划分为 K 个类，则第 i 个样本点对于第 j 个类中心的偏离度表征为(2.2)，其值越小表示第 i 个样本属于第 j 个类的可能性越小，离 j 类越远。

$$S = \sum_{j=1}^{k}\sum_{X_i \in Y_j} P_{ij}\ln p_{ij} \qquad (2.3)$$

P_{ij}：第 i 个样本点对于第 j 个类中心的偏离度

S：类的信息熵值

类的总体信息熵值：设共有 n 个样本点 $X_1, X_2, X_3, \cdots, X_n$ 划分为 K 个类集合，第 j 个类集合用 $Y_j = \{ X_i \mid X_i \in Y_j \}$ 表征，则类的信息熵值表征为(2.3)。

随着类个数由少到多，每个样本点处于各个类边缘的概率加大，他对于各个类中心的偏离度加大，而类的总体信息熵值也变大。在类个数由少到多的过程中，类划分由无序→有序→无序，开始的无序是指划分

太笼统，看不清数据集合的特征，最后的无序是指划分太细碎，对数据集的认识只限于每个小类，没有总体认识。所以我们将整体数据集分为 j 个类的状态定义为数据集的第 j 个状态。第 j 个状态的信息熵值为 Sj。信息熵跳变值：从第 j－1 个状态跳跃到第 j 个状态的信息熵值的改变即 Sj－Sj－1。信息熵跃迁值：从第 j－1 个到第 j 个状态熵值跳变与从第 j 个到第 j＋1 状态熵值跳变的差值即 |(Sj＋1－Sj)－(Sj－Sj－1)| 。我们借用原子物理中的跃迁概念来描述样本数据集在不同类划分状态间的变化，当对样本进行不同类划分时，样本处于不同的状态，在状态间有着信息熵值的微小变化，因此我们从状态变化时信息熵值跃迁的角度来考察类个数的确定情况。在类个数不断增加的过程中，类划分由无序→有序→无序，同时信息熵跃迁值不断变化，而当跃迁值达到最小值时表明从第 j 个到第 j＋1 状态熵值跳变幅度较从第 j－1 个到第 j 个状态熵值跳变幅度在所有跳变幅度中是最小的，也就是数据集整体的不确定程度增加最小，也即信息丰富程度增加最小，那么此时已没有再增加类个数的必要。类个数的变化就此停止，而从第 j－1 个到第 j 个状态变化已达到最佳，我们最终确定 j 个类。

2.2 标 K 均值算法及其在初始类中心确定中的应用

K 均值算法（K－means）是数据挖掘中广泛应用的聚类算法，成功应用于目前的客户市场细分中[6]。该算法以若干个对象作为初始类中心，计算各对象与类中心距离，按照距离最近准则将余下对象逐个归入类中心，作为初始分类。计算各个类的平均值作为新的类中心，然后不断重复迭代，一直到某步归类与前步归类完全一致，则停止运算。其优点在于运算量较小，对于处理样本数据集具有相对可伸缩性和高效性，因此能够快速确定类中心。将运行算法后获得的类中心作为模糊聚类算法的初始类中心可以加快模糊聚类函数的收敛速度，减少迭代步骤，较为快速的得到最终聚类结果。

2.3 模糊 C 均值算法介绍

FCM[7] 设样本 Xi 与 i 类中心 Pi 距离为 $d_{ik}=\sqrt{\|x_k-p_i\|_A}=\sqrt{(x_k-P_i)^TA(x_k-P_i)}$，A 为 S×S 阶的正定矩阵，当 A 取单位矩阵 I 时，此距离对应于欧几里德距离。对于第 i 类的隶属度 u_{ik} 构成软划分隶属度矩阵 $U=[\mu_{ik}]_{c\times n}$，其中 $\sum_{i=1}^{c}\mu_{ik}=1$，其聚类目标是使类内距离总和值为最小，因此聚类目标函数为：

$$\begin{cases}J_m(U,P)=\sum_{k=1}^{n}\sum_{i=1}^{c}(\mu_{ik})^m(d_{ik})^2, m\in[1,\infty]\\ s.t.\ U\in M_{fc}\end{cases}$$

$$M_{fc}=\{U\in R^{cn}\mid \mu_{ik}\in[0,1],\forall i,k;\sum_{i=1}^{c}\mu_{ik}=1,\forall k;0<\sum_{k=1}^{n}\mu_{ik}<n,\forall i\}$$

在使类内距离总和减小的迭代过程中，u_{ik} 与 P_i 不断发生变化，最终使得被划分到同一簇的对象之间相似度最大，而不同簇之间的相似度最小。模糊 C 均值算法是传统 C 均值算法的改进，传统 C 均值算法对于数据的划分是硬性的，而 FCM 则是一种柔性的模糊划分。

3 基于信息熵和 K 均值算法混合迭代的动态模糊聚类算法

由于模糊 C 均值算法（FMC）是在无先验知识的情况下进行聚类，因此其本身无法解决聚类原型初始化参数问题，所以动态模糊聚类算法分为两阶段，第一阶段由信息熵和 K 均值算法混合迭代求出初始类个数以及初始类中心，第二阶段由模糊 C 均值聚类（FMC）求出最终聚类结果。

已知待聚类的样本集合为 $X=\{X_1,X_2,\cdots,X_n\}$，每个样本 X_k 有 s 维属性，由向量来表示即 $x_k=(x_{k1},x_{k2}\cdots,x_{ks}^T$，

算法描述如下：

（1）确定初始聚类个数范围 $[C_{min},C_{max}]$，一般取 $C_{min}=2$，$C_{max}=\sqrt{n}$；

（2）在聚类数目从 $O(n^2c^3s)$ 增加到 C_{max} 的过程中，每对应一个聚类数目 K，利用 K 均值算法求出聚类中心，再计算样本点偏离度的基础上求得信息熵跃迁值

$$S=|(S_{k+1}-S_k)-(S_k-S_{k-1})|;$$

（3）比较 S 值得到在 S 达到最小值 S_{min} 时的聚类个数 c 及此时的聚类中心序列 $P=\{P_1,P_2\cdots,P_c\}$，每个类中心 P_i 表示为 $P_i=\{P_{i1},P_{i2},\cdots,P_{is}\}^T$，作为输出值；

（4）上述类个数 c 及类中心 P 作为初始化参数，设定迭代停止阈值 ε，迭代计数器 step；

（5）根据模糊 C 均值算法推导过程中的 μ_{ik} 公式计算隶属度

$$\forall i,k,if\ \exists d_{ik}^{(step)}>0\Rightarrow\mu_{ik}^{(step)}=\frac{1}{\sum_{r=1}^{c}\left[\frac{d_{ik}^{(step)}}{d_{jk}^{(step)}}\right]^{\frac{2}{m-1}}}$$ 或

$$\exists i,k,d_{ik}^{(step)}=0\Rightarrow\mu_{ik}^{(step)}=1,\mu_{ir}^{(step)}=0(k\neq r);$$

(6) 根据计算得到的隶属度以及模糊 C 均值算法推导过程中的 P_i 公式来更新聚类原型模式

$$p^{(step+1)}=\frac{\sum_{k=1}^{n}(\mu_{ik}^{(step+1)})^{m}x_k}{\sum_{k=1}^{n}(\mu_{ik}^{(step+1)})^{m}}\cdots\cdots i=1,2,\cdots,c;$$

(7) 如果 $\|P^{(step+1)}-P^{(step+1)}\|<\varepsilon$，则聚类算法停止并输出隶属度矩阵 U 和类中心 P，否则 step = step + 1，转向步骤 5，如果超过最大迭代次数也停止迭代；

(8) 根据 λ 不同，在 U 中形成不同的 λ 截集，将样本点划归入不同的类中；

从上述算法中可以看出，在利用信息熵和 K 均值算法进行混合迭代求聚类原型初始化的算法中计算复杂度为 $O(n^{+}s)$，而在模糊聚类算法中计算复杂度为 $O(n^2c^3s)$，因此在第一阶段中如果能很好的确定类个数 C，将明显会降低第二阶段的计算复杂度级别。

4 应用实例

4.1 实例简介

为了验证所提出的基于动态模糊聚类算法的客户细分模型的实际运用效果，本文通过对某省联通 UP 新势力品牌客户 2006 年 8 月的消费数据进行聚类。其中业务类别分为 CDMA 和 GSM，也即 C 网客户和 G 网客户，共有 2532948 条记录，文件大小为 263M。

4.2 客户消费数据预处理

首先利用 ACCESS 数据库进行数据的导入、选择和导出，挑选了 CDMA 客户作为此次聚类分析的对象，利用 SQL 选择语句筛选出 348567 条 CDMA 客户记录。按照表 5.2 中的费用类别将用户的消费记录分为三类：通话类、短信类、数据业务及信息服务类，每类中再分别进行费用额的加总及扣除优惠项目，最后形成 up_08_phone，up_08_short，up_08_info 三张表。部分 SQL 语句如下：

① SELECT ID，业务类别，用户编号，区县编号，地市代码，费用类别，时长，实际通话时长，通话次数，流量，长途计费时长，基本费 + 长途费 + 长附费 + 特服费 + 优惠 fee1 + 优惠 fee2 + 优惠 fee3 + 优惠 fee4 AS 费用 INTO up_sum08 FROM up_200608；

② SELECT 用户编号，地市代码，sum(通话次数) AS 通话次数 1，sum(费用) AS 短消息费 INTO up_08_short FROM up_sum08 WHERE (费用类别 > =140 And 费用类别 < =142) or (费用类别 > =170 And 费用类别 < =180) Or (费用类别 > =1501 And 费用类别 < =1600) Or (费用类别 > =2400 And 费用类别 < =2500) GROUP BY 用户编号，地市代码；

③SELECT 用户编号，地市代码，sum(通话次数) AS 通话次数 1，sum(费用) AS 信息服务及数据业务费 INTO up_08_info FROM up_sum08 WHERE (费用类别 > =181 And 费用类别 < =400) Or (费用类别 > =700 And 费用类别 < =800) or (费用类别 > =1050 And 费用类别 < =1100) Or (费用类别 > =1601 And 费用类别 < =1700) GROUP BY 用户编号，地市代码；

④ DELETE *
FROM up_08_noinfoshort WHERE (费用类别 > =181 And 费用类别 < =400) Or (费用类别 > =700 And 费用类别 < =800) or (费用类别 > =1050 And 费用类别 < =1100) Or (费用类别 > =1601 And 费用类别 < =1700) or (费用类别 > =140 And 费用类别 < =142) or (费用类别 > =170 And 费用类别 < =180) Or (费用类别 > =1501 And 费用类别 < =1600) Or (费用类别 > =2400 And 费用类别 < =2500)；再将三表进行两两关联，按客户 ID 进行分组，SQL 语句在此省略。

因此得到最终的使用 UP 新势力品牌的 CDMA 客户消费数据表 up_08_fee，除去在某一通讯类上消费为 0 的客户，共得到分布在全省 17 个地市的 6776 名客户，再将其导出到文本 unicom.txt 中。

4.3 客户细分过程及其结果

已知 n = 6776，　= 2，≈82，取阈值 10^{-7} 及最大迭代次数 1000，再利用 VC6.0 编程，从文本 unicom.txt 中读取数据，第一阶段进行初始类个数和初始类中心的确定。经过信息熵和 K 均值算法的混合迭代得到的信息熵跃迁值，如图 1 所示。

从图 1 中我们可以发现在跃迁状态为 17，即从(17 个类→18 个类)到(18 个类→19 个类)的跃迁值为最低，表明已没有将 18 个类变为 19 个类的必要。

采用初始类中心中的类中心，取 m = 2，再进行第二阶段的模糊聚类分析。通过聚类运算，得到客户在

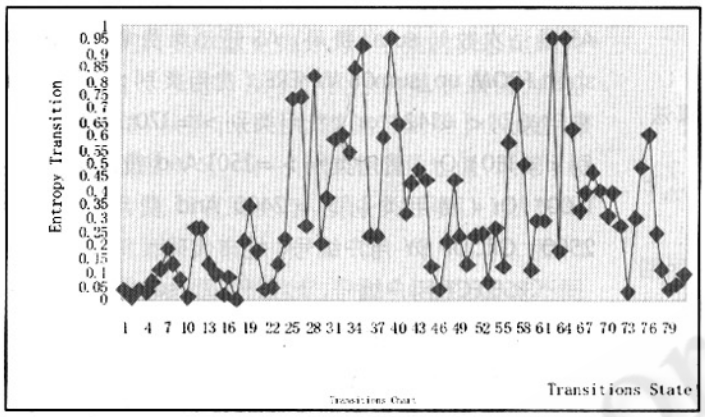

图 1　信息熵值跃迁图

Information Entropy Transition Chart

各类中的分布如图 2 所示。

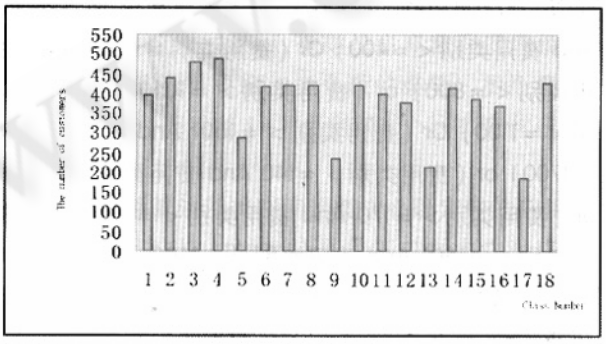

图 2　聚类结果中客户在各类中分布

我们将传统的按距离来确定初始类中心的方法与本文的方法从聚类效果上来对比，从表 4.3 可看到效果较为显著：

表 1　两种聚类方法对比

	传统聚类方法	本文新方法
迭代次数	313 次	234 次
类内距离和	14577.192121	14529.760573
类间距离和	7.276499	7.278561

5　结束语

本文提出了基于信息熵和 K 均值算法混合迭代的动态模糊聚类算法，并将其应用于联通 UP 新势力品牌客户数据，计算所有状态跃迁过程中的信息熵跃迁值，在值达到最小时将类个数和类中心读出，作为模糊聚类的原型初始化参数，通过聚类算法将各指标属性的差别和档次拉开，得到了良好的聚类结果，但仍有大量的工作值的去做：(1) 客户的细分指标仍有进一步分析的必要，随着电信企业推出的业务增多，将客户消费行为与产品相结合来制定细分指标更能反映复杂多变的市场环境。作为对于公司不同营销战略的客户细分指标的再次细化，也有待于进一步研究。(2) 为了优化聚类结果，可以进一步进行测试，得到与实际相吻合的阈值、迭代次数、参数值。(3) 在面对类群的总体营销战略下，对于聚类结果中的每一个类，可针对性地制定相应营销细则，满足特定消费群体的需求。

参考文献

1 Smith, Wendell R. Product Differentiation and Market Segmentation as Alt ernative Marketing Strategies. Journal of Marketing, 1956:3 – 8.

2 Yoram Wind. "Issues and Advances in Segmentation Research" [J]. Journal of Marketing Research. August 1978, 317 – 337.

3 Tsai C. – Y, Chiu C. – C. A purchase – based market segmentation methodology. Expert Systems with Applications, 2004, 27(2):265 – 266.

4 宋华岭等，管理熵理论 – 企业组织管理系统复杂性评价的新尺度，理科学学报，2003，3(16)：19 – 27.

5 L. A . Zadeh. Fuzzy sets. Information and Control, 1965, 8: 338 – 353.

6 林盛、肖旭，基于 RFM 的电信客户市场细分方法，哈尔滨工业大学学报，2006，5(38)：758 – 760.

7 Bezdek J C. Pattern Recognition With Fuzzy Objective Function Algorithms. Plenum Press, New York, 1981.