基于信息熵的蚁群聚类算法在客户细分中的应用①

吴春旭 刘艳泽 苟清龙 (中国科学技术大学 管理学院 安徽 合肥 230026)

摘 要： 传统的蚁群聚类算法需设置较多参数，且聚类时间较长。基于信息熵的蚁群聚类算法通过信息熵改变蚂蚁拾起和放下数据的规则，减少了参数的设置、缩短了聚类的时间，将其应用于客户细分，并且与采用传统的蚁群聚类算法得到的细分结果进行比较分析，实验表明。基于信息熵的蚁群聚类算法可以加快客户细分的聚类进程。

关键词： 信息熵;蚁群算法;聚类;客户细分;应用

Ant Colony Algorithm Based on Entropy for Customer Segmentation

WU Chun-Xu, LIU Yan-Ze, GOU Qing-Long

(Department of Management, University of Science and Technology of China, Hefei 230026, China)

Abstract: The ant-based clustering algorithm needs to set several parameters and cluster for a long time.The ant-based clustering algorithm based on entropy, which uses entropy to amend the ants picking and dropping rules, can reduce the number of parameters and shorten the time of clustering. This paper intends to apply it to the customer segmentation, and compares the segmentation results with the traditional ant-based clustering algorithm. The experiment shows that the ant-based clustering algorithm based on entropy can accelerate customer segmentation.

Keywords: entropy; ant colony algorithm; clustering; customer segmentation; application

1 引言

随着知识经济的迅速发展，市场竞争越来越激烈，产品同质化加剧了企业之间的竞争。如何对客户进行细分，进而针对不同的客户群实施差异化服务和营销，已经成为企业的迫切需求。传统的客户细分方法主要是基于经验的分类或简单的分割[1]。这些方法对一个企业的客户关系管理(CRM)是有一定价值的。随着管理信息系统的广泛应用和电子商务的快速发展，企业已经积累了大量的客户数据，传统的客户细分方法已不能科学有效地处理这些数据，无法满足企业的客户细分需求[2]。

数据挖掘技术可以对客户的所有数据进行智能分类，从中寻找有用的客户行为模式和各种潜在的客户需求，为企业的决策者提供更好的决策支持，已成为CRM 的核心支撑技术之一[3]。基于蚁群算法的聚类算法已经在当前的数据挖掘研究中得到了广泛的应用，该算法不需要预先设定聚类的簇数，而且可以形成任意形状的簇，减少初始参数对聚类结果的影响，但其设置的参数较多，聚类的时间较长。为了减少聚类时间，本文将信息熵和蚁群聚类算法相结合应用于客户细分，利用信息熵减少设置的参数，加快了聚类的速度，从而更好地体现客户特征，帮助企业制定出针对客户的个性化策略,提高客户的价值贡献。

2 蚁群聚类算法

聚类分析是数据挖掘领域中一个重要的研究课题，它根据数据间的相似程度进行分类，使类间的相似性最小化，而类内相似性最大化。经典的聚类分析方法包括分层算法、K 均值算法、模糊 C 均值算法、图论聚类法、神经网络法以及基于统计的方法等。聚

① 基金项目:安徽省自然科学基金(090416240);高等学校优秀青年人才基金(2009SQRS001ZD)
收稿时间:2009-10-30;收到修改稿时间:2009-12-18

类方法应用于客户细分领域，将大量的客户分成不同的类，在每个类里的客户拥有相似的属性，类间的客户属性不同，便于企业确定目标客户，实施产品和服务差异化战略。

蚁群聚类算法的主要思想是[4]：将每一个数据对象随机地投影到平面的一个网格单元中，每只蚂蚁随机地选择网格单元中的一个位置。假设在 t 时刻某只蚂蚁在网格 r 发现一个数据对象 o_i，将对象 o_i 与其邻域对象 oj 的平均相似性定义为(1)式。

$$f(o_i)=\begin{cases}\dfrac{1}{s^2}\displaystyle\sum_{o_i\in Neigh_{s\times s}(r)}\left[1-\dfrac{d(o_i,o_j)}{\alpha}\right], & 若 f(o_i)>0\\ 0, & 其他\end{cases} \tag{1}$$

其中，$Neigh_{s\times s}(r)$ 表示 r 周围的以 s 为边长的正方形区域；d(oi, oj)表示对象 oi 和 oj 在属性空间中的距离，通常采用欧式距离；α 为调节数据对象间相似性的参数，它决定了聚类数目和收敛速度。

蚂蚁在运动过程中拾起对象 o_i 的概率 $p_p(o_i)$和放下对象的概率 $p_d(o_i)$分别是：

$$p_p(o_i)=\left(\frac{k_1}{k_1+f(o_i)}\right)^2 \tag{2}$$

$$p_d(o_i)=\begin{cases}2f(o_i), & 若 f(o_i)<k_2\\ 1, & 若 f(o_i)\geq k_2\end{cases} \tag{3}$$

其中 k_1，k_2 都为阈值常量。

蚂蚁沿着网格单元移动时，根据公式(1)计算对象的平均相似性。如果蚂蚁未负载，则根据公式(2)计算“拾起”的概率 p_p，若 p_p 大于某一个随机概率，而同时该对象未被其他蚂蚁“拾起”，则蚂蚁“拾起”该对象，随机移往别处，并标记自己有负载，否则，蚂蚁拒绝“拾起”该对象，而随机选择其他对象。如果蚂蚁是有负载状态，则根据公式(3)计算“放下”概率 p_d，若 p_d 大于某一个随机概率，则蚂蚁“放下”该对象，并标记自己未负载，而重新选择一个新的数据对象。

蚁群聚类算法能实现完全分布式控制，并具有自组织性、可扩展性、健壮性等特征，而且采用蚁群模型进行聚类更加接近实际的聚类问题[5]。但是蚁群聚类算法需要设置的参数多，参数设置不当就会导致聚类结果不理想。而拾起与放下的规则会导致一个数据对象被多次拾起或放下，从而影响聚类的进程。

3　基于信息熵的蚁群聚类算法

针对蚁群聚类算法设置参数多、计算时间长的缺点，将信息熵理论用于蚁群聚类算法，通过信息熵的计算与比较，改变了拾起和放下数据的规则，减少了设置的参数，提高了聚类的性能。

3.1 信息熵定义

Shannon 提出的信息熵定义为：假设是一个随机变量， X 是其可能的取值集合（连续型数据需离散化），p(x)是取 x 值的可能性函数，信息熵 E(x)定义为：

$$E(x)=-\sum_{x\in X}p(x)\log p(x) \tag{4}$$

一个多变量向量 $x=\{x_1,x_2,\cdots,x_n\}$的信息熵为：

$$E(x)=-\sum_{x_1\in X_1}\cdots\sum_{x_n\in X_n}p(x_1,x_2,\cdots,x_n)\log p(x_1,x_2,\cdots,x_n) \tag{5}$$

其中 $p(x)=p\{x_1,x_2,\cdots,x_n\}$是多变量可能分布函数，$X_1,X_2,\cdots,X_n$ 是相应向量项的可能取值集合(连续型数据需离散化)。

3.2 算法思想

于信息熵的聚类算法均依据这样一个事实：包含聚的子空间的信息熵比不包含聚的信息熵小[6]。借鉴这一思想，可以将信息熵引入蚁群聚类算法，改变蚂蚁拾起与放下数据对象的规则，算法的主要是思想是[7]：

一个未负载的蚂蚁移到对象 o_i 之处，计算周围 s×s 区域中对象的信息熵，假设未拾到对象 o_i 之前的信息熵为 E_1，拾起对象 o_i 之后该区域的信息熵为 E_2，那么拾起的规则为：若 $E_1>E_2$，则拾起对象 o_i。

一个负载对象 o_i 的蚂蚁移到空白之处，计算周围 s×s 区域中对象的信息熵，假设未放下对象 o_i 之前的信息熵为 E_1，放下对象 o_i 之后该区域的信息熵为 E_2，那么放下的规则为：若 $E_1>E_2$，则放下对象 o_i。

假设每一个对象包含 n 个互为独立的属性 $A_1,A_2,\cdots,A_n$，各属性的可能取值集合为 $x_1,x_2,\cdots,x_n$，s×s 区域中对象的信息熵可以按以下公式计算：

$$E(s^2)=-\sum_{i=1}\sum_{x\in X_n}p(x)\log p(x) \tag{6}$$

$$p(x)=\frac{x}{Y} \tag{7}$$

其中 x 为 s×s 区域中 $A_i=x$ 的对象个数；Y 为 s×s 区域中对象的总数。

基于信息熵的蚁群聚类算法只需要设置 3 个参数蚂蚁数目 N_{ant}、蚂蚁迭代次数 t_{max} 和 s，影响拾起或者放下对象的因素只有参数 s。每次放下对象能减少小块区域的信息熵，每次拾起对象能增加小块区域的信息熵，根据包含聚的子空间的信息熵比不包含聚的信息熵小的思想，使得同类型的数据对象能够较快的聚集在一起，但是产生的结果是局部最优。通过调整 s 的大小，可以减少小块聚的产生。

4　基于信息熵的蚁群聚类的客户细分

为了验证基于信息熵的蚁群聚类算法在客户细分中的应用效果，本文对某地区的移动用户在 2008 年 5 月至 7 月的消费数据进行聚类。该地区的移动用户总数 438018。通过使用数据库技术对数据进行导入、选择和导出等预处理，得到 7893 条客户记录作为此次客户细分的研究对象。

4.1 客户数据的预处理

首先利用 ACCESS 数据库进行数据的导入、选择和导出等预处理，利用 SQL 选择语句筛选出 7893 条客户记录作为本次聚类分析的研究对象。部分 SQL 语句如下。

① SELECT 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，长途计费时长，基本费+长途费+长途费+特服费+优惠 fee1+优惠 fee2+优惠 fee3+优惠 fee4 AS 费用 INTO g_sum09 FROM g091;

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 g_09_short FROM g_sum09 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;

③ SELECT 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 g_09_info FROM g_sum09 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 SELECT 用户编号，地市代码，sum(通话次数) AS 通话次数 1，sum(费用) AS 通话费 INTO g_09_phone FROM g_09_phone1 GROUP BY 用户编号，地市代码;

4.2 客户细分过程

① 初始化。将 7893 个客户数据对象随机的分布在 200×200 的二维网格中，每一个数据对象随机地投影到一个网格单元中。用 20 只蚂蚁在网格中独立的移动并完成聚类，蚂蚁的迭代次数为 50000。整个聚类过程采用 6×6 的网格区域作为局部面积；

② 每一只蚂蚁随机的选择网格中的一个位置，并且初始状态为未负载；

③ 对于每一个蚂蚁，如果蚂蚁未负载，且在位置 r 处有客户数据对象 o_i，那么首先计算周围 6×6 区域中未拾起对象 o_i 时的 p(x)。在对客户属性进行判断时，如果需要分类的客户属性 $A_1,A_2,\cdots A_n$ 为不连续的值，对于两个客户数据对象 o_j 和 o_k，若 $A_j=A_k$，则对象的两个属性为同一；如果需要分类的客户属性 $A_1,A_2,\cdots A_n$ 为连续的值，对于两个客户数据对象 o_j 和 o_k，若 $|A_j-A_k|<\varepsilon$，则对象的两个属性为同一，其中为该属性范围内的允许误差，ε =0.1；

④ 计算周围 6×6 区域中未拾起客户数据对象 oi 时的信息熵 E1；

⑤ 计算当蚂蚁拾起客户数据对象 o_i 时周围 6×6 区域中的 p(x)，对客户属性的判断同步骤③；

⑥ 计算拾起客户数据对象 o_i 后该区域的信息熵 E_2，同时与 E_1 进行比较，若 $E_1>E_2$，则蚂蚁拾起客户数据对象 o_i；

⑦ 若蚂蚁负载客户数据对象 o_i，并且所在之处为空，则计算周围 6×6 区域中未放下客户数据对象 o_i 的 p(x)，对客户属性的判断同步骤③；

⑧ 计算蚂蚁未放下客户数据对象 o_i 时该区域的信息熵；

⑨ 计算蚂蚁放下客户数据对象 o_i 后周围 6×6 区域中的 p(x)，对客户属性的判断同步骤③；

⑩ 计算蚂蚁放下客户数据对象 o_i 后该区域的信息熵 $\overline{E}_2$，同时与 $\overline{E}_1$ 进行比较，若 $\overline{E}_1>\overline{E}_2$，则放下客户数据对象 o_i，蚂蚁再继续移动；

⑪ 通过以上过程最终将 7893 条客户记录聚类为 10 个客户群。

4.3 客户细分结果分析

通过应用基于信息熵的蚁群聚类算法，将 7893 个客户最终分为 10 类(见图 1)，其中每一类客户具有相似的特征，类间的客户特征差别较显著。

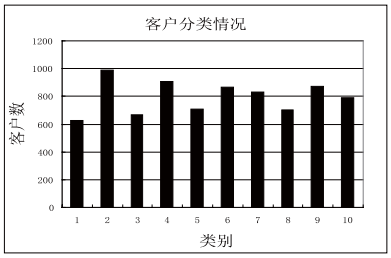

图 1　客户分类图

通过对以上不同客户群的消费特征进行分析，可以得出每类客户群的所属客户类型(见表 1)。

表 1　各类客户群所属客户类型

类别	所属客户类型	类别	所属客户类型
1	高价值型	6	商旅型高价值型
2	大众办公型	7	高忠诚度型
3	低价值无通话型	8	本地高价值型
4	漫游型	9	一般价值客户
5	高忠诚大众办公型	10	大众夜话型

为了进一步验证基于信息熵的蚁群聚类算法在客户细分中的有效性，本文同时将蚁群聚类算法应用于客户细分，细分结果如图 2 所示。

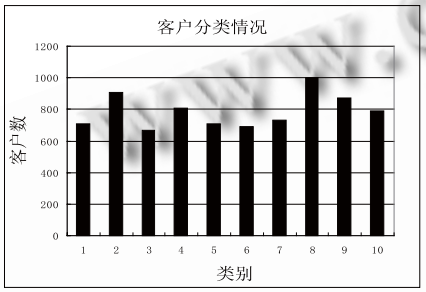

图 2　客户分类图

从误分类错误率和迭代次数上对两算法的性能进行比较，结果见表 2，可以看出两算法在客户细分方面存在明显的差别。蚁群聚类算法经过 100000 次循环后可以比较准确地对客户数据聚类，但是循环次数较多，收敛速度较慢；而基于信息熵的蚁群聚类算法在每一次对数据对象拾起或者放下时，都不依赖于概率而是通过信息熵直接进行对应的动作，因此能够经过较少次的循环较快的收敛，但是由于局部解的存在，即聚类过程中会形成多个小类，而蚂蚁很难将这些小类中的对象拾起并和其他类进行组合，所以在误分类错误率上高于蚁群聚类算法。

表 2　两种聚类方法对比

	蚁群聚类算法	基于信息熵的蚁群聚类算法
误分类错误率	6.3%	11.6%
迭代次数	100000	50000

基于信息熵的蚁群聚类算法只是减少了主循环的次数，并没有缩小算法的时间复杂度，因此时间复杂度与蚁群聚类算法相同，均为 O(tmaxNant)。

5　结语

该算法通过信息熵改变了蚁群聚类算法中拾起和放下数据对象的规则，减少了设置的参数，加快了聚类的速度，得到良好的聚类效果，从而有效的实现客户细分，为企业制定个性化的服务和营销提供依据。但是仍有大量的内容需要进一步研究：①随着企业推出的业务的增多，需要将客户的细分指标进一步的细化，以确保能准确地反映市场环境；②为了优化聚类结果，可以进一步研究聚类过程，提高聚类的准确率；③对于聚类结果中的每一类客户，可以根据其特征制定出相应的营销策略，满足特定客户群的需求。

参考文献

1 Zhang GZ, Gao J. The Customer Segmenta tion Study Based on Data Mining of CRM. Journal of Marketing, 2005:24－25.

2 Lai X. An segmentation study on enterprise custo mers based on data mining technology. Jianxun Xhne, Zhengbing Hu. 2009 First International Work shop on Database Technology and Applications, DBTA. Wuhan: IEEE, 2009:247－250.

3 Kim W. On Business Intelligence System. Proc. of 2nd International Conference on Worldwide Computing and Its Applications. Japan:1998.337－348.

4 段海滨.蚁群算法原理及其应用.北京:科学出版社, 2005.291－292.

5 郭会林,苏一丹.一种基于混合策略的蚁群聚类算法. 计算机工程与应用, 2008,44(6):154.

6 Cheng CH, Fu AW, Zhang Y. Entropy-Based subspace clustering for mining numerical data. Proc. of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 1999.84－93.

7 刘波.一种利用信息熵的群体智能聚类算法.计算机工程与应用, 2004,35:180－183.