基于SOM的市场细分研究

吕 昱 程代杰
(重庆大学计算机学院 重庆 400044)

摘 要 本文应用数据挖掘技术对市场细分的研究包括:如何利用SOM聚类技术解决市场细分问题;如何将SOM聚类的市场细分结构结果可视地呈现给市场决策人员的问题。这两部分研究整合为企业市场战略集成一个研究途径,包括:偏好数据的收集与前处理步骤、偏好数据聚类步骤、偏好数据的可视化步骤。实验结果表明,本文提出的研究途径成功发现了人工数据集中预设的聚类模式,与通常研究途径相比具有明显的优点。在实际市场数据分析中,获得了与事实相符的结论并提供了有价值的决策支持信息。

关键词 偏好序列,聚类,市场细分,自组织特征映射

SOM-Based Market Segmentation Research

LU Yu CHENG Dai-Jie
(College of Computer Science, Chongqing University ,Chongqing 400044)

Abstract The research in market segmentation includes two main parts. We focus firstly on discussing the market segmentation problem by applying SOM clustering technique in data mining discipline. The second part is focus on displaying market segmentation structure. We apply visualization technique to represent the market structure clearly in a two dimensional plane so that the marketers can make their market strategies easier. The two main parts are organized as an integrated approach. Such an approach includes three core steps: preference data collecting step, preference data clustering step by SOM neural networks and visualization step by ideal point model. The experiments show that the approach yields meaningful results and is comparable and complementary to the most general ones.

Keywords Preference order, Cluster analysis, Market segmentation, SOM

1 引言

市场细分由Wendell Smith在1956年[1]提出,并成为市场营销领域的一个重要概念。理解市场结构和准确预见消费者行为,对处于激烈竞争环境的公司合理配置资源和发现潜在市场机会具有重要意义。一类市场细分的途径是利用某些先验因素,比如地理区域、职业、生活方式等,直接对消费者进行市场细分。另一类市场细分途径利用市场调查数据,通过聚类分析手段,对市场进行后验细分[3]。由于后验细分不需要对特定市场的先验知识,且能发现市场调研数据中内含的知识,因此被认为是更好的市场细分手段[2]。本文沿后一途径,提出利用偏好数据进行市场细分,并对细分结果进行直观呈现的完整的解决途径。论文第2节给出基于偏好数据市场细分问题的形式化表述;第3节给出基于偏好数据市场细分和可视化问题的实现算法;第4节是应用本文的研究途径,对人工和真实数据集进行实验结果分析,最后是全文总结。

2 问题描述

2.1 偏好数据聚类问题

设E表示评估者集合:$E=\{e_1,e_2,\cdots,e_n\}$;A表示被评价对象的集合:$A=\{a_1,a_2,\cdots,a_p\}$。

定义1 偏好序列是评估者对被评价对象依据个人偏好给出的全序或偏序排列。

全序:对集合A中所有被评价对象给出完整排序;

固定偏序:从集合A中选择固定数目的评价对象,按偏好给出排序;

可变偏序:从集合A中选择固定数目的评价对象,按偏好给出排序;

本文仅讲座全序情况。

对每个e_i给出的偏好序列$A_i=(a_{i_1}a_{i_2}\cdots a_{i_p})$,定义坐标算子:

$\rho_i:A\to\{1,2,\cdots,p\}$,$\rho_i(a_j)=r_j$,$a_j\in A$,$r_j$是$a_j$在$A_i$中的位置序号。

对坐标算子还可以定义其逆算子:

$\rho_i^{-1}:\{1,2,\cdots,p\}\to A$,$\rho_i^{-1}(r_j)=a_j$,$\forall r_j\in\{1,2,\cdots,p\}$,$a_j$是在$A_i$中$r_j$位置对应的元素。

定义2 $\forall e_i,e_j\in E$,定义e_i到e_j的距离$d^{(E)}(e_i,e_j)=\sum_{a\in A}|\rho_i(a)-\rho_i(\rho_j^{-1}(\rho_i(a)))|$。

定理1 设$e_1,e_2\in E$,且e_1,e_2偏好排序互为逆序,则$d^{(E)}(e_1,e_2)=\max_{e_i,e_j\in E}(d^{(E)}(e_i,e_j))$。

定理2 $\forall e_i,e_j\in E$,$d^{(E)}(e_i,e_j)=d^{(E)}(e_i,e_j)$。

由定理1、2和定义1,当两个评估者给出排序一致时,其距离为0;当他们给出排序相反,则距离为最大值,这与直觉吻合,同时该距离定义满足正定、对称性,并可证明满足三角不等式,因此定义1作为描述评估者相似性的距离度量是较

吕 昱 博士研究生,主要研究方向为聚类技术在产品研发早期阶段的应用;程代杰 教授,博士生导师,研究领域为并行处理,网络技术,商业智能等。

合理的。在定义1的基础上，将消费者对产品的偏好排序转化为其相互之间的相似(异)性矩阵，便可利用合适的聚类算法进行聚类操作实现市场细分。限于篇幅，定理证明略。

2.2 聚类结果可视化问题

在市场营销领域，James Lattin 等学者[7]将偏好数据可视化问题称为偏好的多维尺度分析。通常有两种技术用于将偏好数据映射为直观的二维图形。Lattin 对比了非度量多维尺度分析模型和度量多维尺度分析模型两种技术，指出了两种模型的各自优缺点。本文采用非度量多维尺度分析模型，给出适当的相似性定义，避免了 Lattin 所提的相似性矩阵中大部分元素无定义的缺点。

定义3 $\forall e_i \in E, a_j \in A$，定义 e_i 到 a_j 的距离为：

$$d^{(EA)}(e_i, a_j) = d^{(EA)}(a_i, e_i) = \alpha \cdot \rho_i(a_j)$$

$\alpha \in R^+$，为距离调节系数。

定义4 设 $U = E \cup A = \{u_1, u_2, \cdots, u_{k+p}\}$，$\forall u_i, u_j \in U$，定义 u_i 和 u_j 的相异性为：

$$dissim(u_i, u_j) = \delta_{ij} =$$

$$\begin{cases} d^{(EA)}(u_i, u_j); & u_i \in E, u_j \in A \text{ 或 } u_i \in A, u_j \in E \\ d^{(E)}(u_i, u_j); & u_i \in E, u_j \in E \\ \sigma_{ij}; & u_i \in A, u_j \in A \end{cases}$$

按定义4，可视化问题转化为寻求二维平面上一组点使其形成的欧式距离矩阵与由定义4得出的相异性矩阵尽量吻合的问题。为度量两矩阵吻合程度，可视化问题进一步转化为一个标准的非线性优化问题。

$$\min \text{STRESS} = \sum_{i=1}^{p+1} \sum_{j=i+1}^{p+k} \left| \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} - \delta_{ij} \right| \quad (1)$$

$$s.t.\ x_i, x_j \in R。$$

3 实现算法

3.1 聚类步骤

由定义1，偏好数据聚类问题转化为标准的聚类问题。通过研究聚类算法相关文献[8]，本文最终选择自组织特征映射(SOM)聚类算法。一方面，SOM 具有很好的性能，能处理很大的数据集；另一方面，不像 K-methods 类算法，要先给出 K 值，而且 SOM 对孤立点不敏感。因此 SOM 特别适合市场细分问题。

输入：所有被调查对 p 种产品的偏好序列。
输出：训练后的神经网络。
1. 将所有偏好序列转移成偏好向量集，按定义1，偏好向量的分量为偏好序列对应符号的坐标算子；
2. 初始化 SOM；
3. 若没达到设定的网络训练次数，则随机选择偏好向量样本作为输入；否则转 7；
4. 计算输出层最匹配神经元 BMU；
5. 更新输出层获胜节点的神经元和该神经元某邻域范围神经元的节点权重；
6. 调整 SOM 算法收敛参数，转到 3；
7. 将偏好向量集的每个向量，映射到已经稳定的 SOM 网络，输出层神经元记录其被命中的次数。

3.2 可视化步骤

按定义2和定义4，应用非度量多维尺度模型。聚类步骤产生结果的可视化问题被转化为标准的非线性最优化问题。本文比较研究了最速下降法、模拟退火法和希尔爬山数值算法。实验结果表明，模拟退火法往往发现更好的解。最速下降法效率更高。

3.3 时间复杂度分析

聚类步骤算法包括3个部分：偏好数据转化为偏好向量；训练 SOM 网络；将偏好向量集的每个向量映射到已经稳定的 SOM 网络。总的算法复杂度大致为 $O(C \cdot n \cdot d) + O(p \cdot n) + O(n \cdot d)$。对给定的 SOM 网络和被评价产品集，聚类步骤的算法花费时间与偏好数据集大小基本是线性关系，时间复杂度可以表示为 $O(K_1 \cdot n)$，K_1 是常数。

可视化步骤时间花费集中在算法的第4步和第5步。时间复杂度可表示为 $O(K_2 \cdot n_v^2)$，K_2 是常数，n_v 是 SOM 输出层权重向量的数量。

4 实验及分析

为评估本文提出的市场细分研究途径，我们使用了人工和真实的数据集，实验数据包括用程序随机产生的人工偏好数据集(按预设规则随机产生偏好序列)和企业通过专业市场公司调查获得的真实数据。实验运行环境是 Pentium III 550 MHz PC，内存为 768MB。操作系统是 Windows2000 Professional。所有算法在 Matlab6.5 R13 环境实现。

人工偏好数据集上的实验：假定偏好数据集是针对9种被评价产品给出的偏好排序，即被评价产品集合为 $A = \{a, b, c, d, e, f, g, h, i\}$。并假定评估者有3类，设为 $A^{(1)}$，$A^{(2)}$ 和 $A^{(3)}$。$A^{(1)}$ 类评估者最喜欢 a 或 b，不喜欢 h 和 i；$A^{(2)}$ 类评估者比较喜爱 c、d、e、f 及 g，但不喜爱 a 和 b；$A^{(3)}$ 类评估者非常喜爱 h 和 i。$A^{(1)}$、$A^{(2)}$ 和 $A^{(3)}$ 类包含的评估者数量分别为100，60，130。按照这一规则，我们随机生成了4个偏好数据集进行实验，结果见图1。

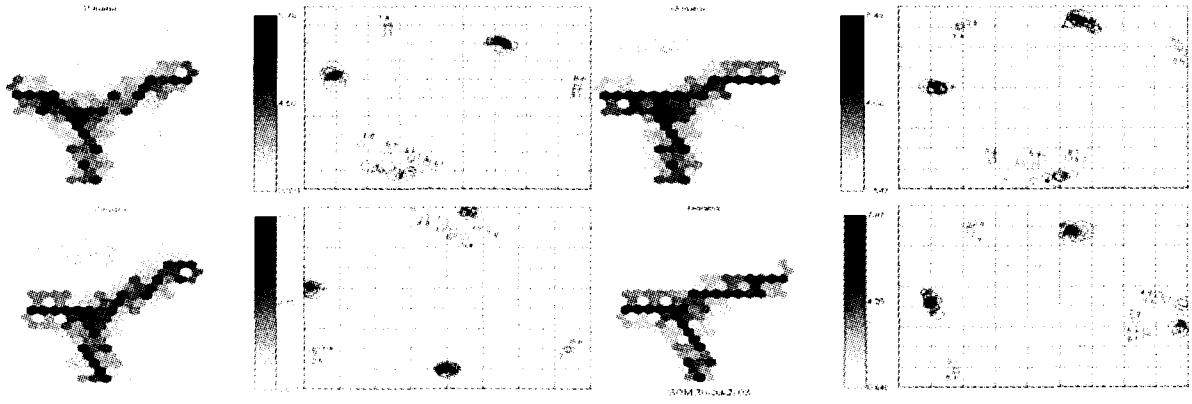

图1 第一组实验：预设3个聚类的4个偏好数据集应用聚类步骤的结果

真实数据集上进行的实验：本部分数据来自国内某知名汽车公司对其新车型构成竞争的15种现有车型进行市场调查的数据。汽车公司最初只提供了400个评估者对15种车型的偏好数据。应用本文方法、实验结果见图2。由于分析

结果明显支持存在4个细分市场，我们向公司求证该结论。汽车公司相关负责人员证实，调查数据来自国内4个主要的汽车消费城市。后续研究证实，4个细分市场的确对应4个城市的被调查者。最终，本文研究成果得到企业充分认可。

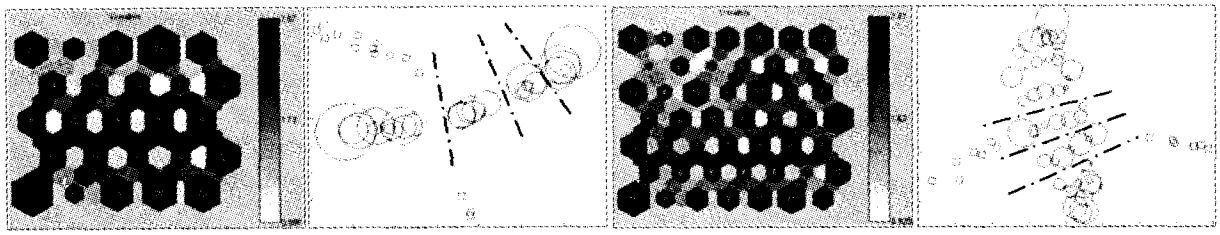

图2 真实数据集算法输出结果

结论 本文提出利用偏好数据进行市场细分，并对细分结果进行直观呈现的完整解决途径。本文对该途径给出了完善的数学模型和实现算法，并经受了实际数据集的检验，获得非常好的评价。本文提出的研究途径细分结果稳定，可以控制市场细分的粒度，具有良好的应用特性。

参考文献

1 Smith W R. Product differentiation and market segmentation as an alternative marketing strategy. Journal of Marketing, 1956, 21:3～8
2 Neal W D. Advances in market segmentation. Marketing Research, 2001, 13(1): 14～18
3 Hsu Tsuen-Ho, et al. The fuzzy clustering on market segment. In: The ninth IEEE international conference on Fuzzy Systems, 2000. 621～626
4 CS Inc and MB Research & Consulting. Market Segmentation for Ridership Forecasting. www. watertransit. org
5 Hoek J, et al. Market segmentation - A search for the Holy Grail? Journal of Marketing Practice: Applied Marketing Science, 1996, 2(1): 25～34
6 MacEwan, et al. Preference mapping: a review. Campden & Chorleywood Food Research Association Review, 1998(6)
7 Lattin J, et al. Analyzing Multivariate Data. 2003. 244～255
8 Fasulo D. An Analysis of Recent Work on Clustering Algorithms. Scientific Literature Digital Library, citeseer. nj. nec. com. 1999

(上接第97页)

solution)作为备选解。再调用行为 choose 从备选解中选出一个解向买方 Agent 推荐。

4)30～34行，若卖方 Agent 拒绝接受建议，卖方 Agent 会继续调用行为 choose 从备选解中选出另外的向 Agent 推荐，显然，若买方 Agent 一直拒绝，那么最后一个向买方 Agent 推荐的解就是 last_solution，也就是说，买卖双方至少能在 last_solution 上达成交易。

5)行为 receive 和 propose 分别负责接收买方 offer 和向买方发送 offer。

买方 Agent 的协商策略包含两方面，一方面是在效用上的权衡策略，在协商开始时先调用行为 *get_threshold* 确定效用门限，效用门限一旦确定，在协商过程中就不会做出让步。当获得一个可接受的解后，卖方 Agent 可能会请求买方 Agent 考虑另外的解，若卖方 Agent 提供的其他令买方获得更高的效用，则接受，否则维持原解。另一方面是在对产品属性的约束上的最小让步策略。对产品特性的控制是由 *cut_level* 实现的。*cut_level* 是一个向量，向量中的每个值是当前买方对一个模糊约束满足度的要求。协商开始时，*cut_level*=1，买方 Agent 争取获得在产品特性上的最高得分，也就是要求卖方提供使所有模糊约束满足度都达到1的产品。所谓让步就是降低对某个模糊约束满足度的要求，行为 relax 保证每次的降低都是在产品特性上的必要的最小程度的让步。

卖方 Agent 的协商策略可分为两个阶段：第一阶段卖方 Agent 总是先根据买方 Agent 的要求寻找合适的产品，在所有合适的产品中又总是优先选出令自己获利最高的作为提议，当找不到满足要求的商品时，它会试图要求买方 Agent 放松约束；当买方 Agent 接受一个解之后，卖方 Agent 会试图在能令自己获得更高利润的产品中选择可能能够改进买方 Agent 效用的解(产品特性虽不如当前解，但价格也更低)，要求买方 Agent 重新考虑，直到最终双方都确定某个解。

买卖双方的协商策略能够保证：1)若存在可能的解，则一定能达成交易；2)获得的解是 Pareto 最优的。

结论和进一步的工作 在本文中，我们基于对零售市场特征及商家和消费者协商行为的分析，提出了一个综合采用模糊约束思想和多属性效用理论的零售电子市场商品交易自动协商模型，给出了买方 Agent 和卖方 Agent 的形式化模型，介绍了它们各自的行为协议和采用的协商策略。协商策略的设计保证了获得的解是 Pareto 最优的，这也就意味着买卖双方达成了一个较为公平的交易，符合实质利益协商法的原则。当然，模型中还存在一些需要进一步改进的地方。首先，买方 Agent 效用门限的确定对协商结果的影响还需进一步的实验评估；第二，买方 Agent 在效用上采用了权衡策略，进一步的研究可以考虑加入时间约束，在效用上采用随时间让步的策略，同时，约束放松策略也应该相应作出调整；第三，可以为卖方 Agent 设计更复杂的协商策略，比如它是否可以为了获得更高的利润而在有解的情况下故意要求买方 Agent 放松约束。

参考文献

1 TechWeb. com. U. S. Online Retail Sales Expected to Double In Six Years. http://www. techweb. com/wire/30000066, 2004-08-23
2 Guttman R, Maes P. Cooperative vs. Competitive Multi-Agent Negotiations in Retail Electronic Commerce. In: Proc. of the Second Intl. Workshop on Cooperative Information Agents (CIA'98). Paris, France: July 1998
3 He Minghua, Jennings N R, Leung Ho-Fung. On Agent-Mediated Electronic Commerce. Knowledge and Data Engineering, 2003, 15(4): 985～1003
4 Zeng D, Sycara K. Bayesian Learning in Negotiation. International Journal of Human-computer Studies, 1998, 48(1): 125～141
5 Faratin P. Automated Service Negotiation Between Autonomous Computational Agents: [PhD thesis]. University of London, Department of Electronic Engineer Queen Mary & Westfield College, Dec. 2000
6 Luo X, Jennings N R, Shadbolt N, et al. A Fuzzy Constraint Based Model for Bilateral Multi-issue Negotiations in Semi-competitive Environments. Artificial Intelligence Journal, 2003, 148(1-2): 53～102
7 Schiffman L G, Kanuk L L. 消费者行为学(第七版). 上海：华东师范大学出版社，2002
8 Willmott S, Calisti M, Faltings B, et al. CCL: Expres_sions of Choice in Agent Communication. In: Proc. of the Fourth Intl. Conf. on Multi-Agent Systems, 2000. 325～332