第 36 卷 第 4 期 计 算 机 工 程 2010 年 2 月
Vol.36 No.4 Computer Engineering February 2010
·开发研究与设计技术· 文章编号：1000—3428(2010)04—0276—03 文献标识码：A 中图分类号：TP311

基于数据挖掘的供应链产品优化配置

初佃辉，郑宏珍
(哈尔滨工业大学(威海)企业智能与服务计算中心，威海 264209)

摘 要：针对供应链的特点，以客户偏好序列数据为切入点，提出客户偏好取向与客户特征属性间的关联关系模型，借鉴数据挖掘的符号序列聚类方法，研究符号类型序列数据对应的性质，从形式化和实例化 2 个方向讨论符号序列相似性问题，对偏好符号序列聚类问题的本质进行分析，研究如何应用自组织特征映射作为符号序列的聚类算法，并对聚类模型进行比较，使得从消费者偏好进行市场细分结构研究的研究途径在实际应用中得以实现。

关键词：供应链；数据挖掘；优化配置；符号序列聚类

Optimal Configurations of Supply Chain Products Based on Data Mining

CHU Dian-hui, ZHENG Hong-zhen
(Corporation Intelligent and Service Computing Center, Harbin Institute of Technology(Weihai), Weihai 264209)

【Abstract】 Aiming at the characteristics of supply chain, considering the customer preferences sequence data, this paper presents an association relation model relating customer preferences and customer characteristic attributes. Referencing cluster method for the symbol sequence in data mining, corresponding properties of the data from type symbol sequence are studied. The research follows two directions, formal and instantiation, to discuss similarity issues of symbol sequence in which essential problems in cluster from preference symbol sequence clustering are analized. This paper studies on how to apply self-organizing feature map as a symbol of the sequence clustering algorithm, and compares clustering model, thus enabling from the consumer preference for market segmentation studies in the structure means in practical application is realized.

【Key words】 supply chain; data mining; optimal configurations; symbolic sequence clustering

1 概述

随着市场竞争的日趋激烈和先进技术在现代工业中的广泛应用，供应链企业之间的竞争开始转向基于时间和基于客户需求的竞争。高质量、低成本、短交货期、个性化产品已经成为现代企业追求新的竞争优势的一种必然趋势。采取何种生产方式以适应瞬息万变和日益个性化的市场环境是制造业面临的一个重要课题。

在市场上，一个有竞争力的新产品诞生，要经历产品策划、概念设计、详细设计、工艺/工装设计、样机试制/试验、设计定型、投产准备、批量生产、销售和售后服务等生命周期的各个阶段。在产品生命周期的各个阶段中，对新产品功能、性能、成本和顾客满意度的影响最大的是早期的产品策划，因为它一端联系着顾客对未来新产品的潜在需求(需求偏好)及当前市场竞争对手的状态，另一端联系着新产品的概念定义，即将顾客的潜在需求和需求偏好转换为目标新产品的功能技术特性，新产品的后续概念设计、详细设计到生产、销售的全过程都是依据产品策划而进行的。因此，产品策划是否准确地把握了未来市场和顾客的需求，理清自己与竞争对手在市场所处的格局，就直接关系着后续新产品设计和生产销售的成败，可以说产品策划是影响未来产品竞争力和企业竞争力最重要的环节。

在当前国际国内竞争日益加剧的环境下，消费品供应链制造企业如何正确认识市场细分结构，协助企业解决市场细分结构分析和研究消费者偏好和特征，以便于科学部署与实施其新产品结构和销售网点，从而确定有效的物流网络，对制造行业这一国民经济支柱产业也具有积极的意义。

为解决上述问题，文献[1-2]利用统计学传统聚类方法得到市场细分结构，选择消费者的职业、收入、年龄、性别等特征数据作为细分变量。在实际应用中，不同的细分变量会导致不同的市场细分结果[3-4]。

为此，本文提出以偏好序列为切入点，研究市场细分结构，并进一步发现消费者特征与其不同偏好行为内在的关联关系，从而指导企业在新产品概念设计阶段，实施正确的市场定位及部署相关的产品战略。

2 供应链产品符号序列聚类

在供应链的管理中，市场细分目前已经成为供应链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。

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。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

基金项目：国家“863”计划基金资助项目(2008AA04Z101)；山东省科技攻关基金资助重大项目(2008GG10004010)；山东省自然科学基金资助重点项目(2007ZRA1000)
作者简介：初佃辉(1969－)，男，副教授、硕士，主研方向：数据挖掘，智能计算；郑宏珍，教授、博士
收稿日期：2009-11-12 **E-mail：**hithongzhen@163.com

细分手段[5]。

本文的研究方法是从消费者对多种评价对象的心理趋向是否存在聚类这个角度出发，研究基于聚类的市场结构分析方法。

定义 1 设 $\Sigma=\{a_1,a_2,\cdots,a_k\}$ 为有限符号表，其中的 l 个符号 $\{a_1,a_2,\cdots,a_l\}$ 构成的有序集称为符号序列，记为 $s=<a_1,a_2,\cdots,a_l>$，并称 l 是 s 的长度，记为 $|s|$，空符号序列记为 $\varepsilon,|\varepsilon|=0$。$\Sigma$ 上所有有限长度符号序列集合记为 Σ^*。

例如，设 $\Sigma=\{1, 2, 3, 4, 5, 6, 7, 8, 9, A, B, C, D, E, F\}$表示 15 种车型，是对客户偏好排序数据经过处理后得到符号序列。

定义 2 设 $P=\{S_1,S_2,\cdots,S_n\}$，S_t 是 Σ^* 上的某个符号序列。符号序列聚类是指寻找 P 上的划分 $P_1,P_2,\cdots,P_k$，使属于同一划分的符号序列间的相似性尽量大，而属于不同划分的符号序列间相似性尽量小。

定义 3 $s_1,s_2\in S,\tilde{S}_{12}$ 是 s_1,s_2 产品子序列集合，设 $\tilde{s}_z\in\tilde{S}_{12}$，若 $\forall\tilde{s}\in\tilde{S}_{12}$ 满足 $|\tilde{s}_z|\geq|\tilde{s}|$，则称 $\tilde{s}_z$ 是 s_1,s_2 的最大公共产品子序列(Largest Public Sequences, LSP)，记为 $LPS(s_1,s_2)$。

为了对供应链产品符号序列进行分析，本节给出供应链产品符合序列的形式化描述。

目前已经建立了许多相似性度量模型，深刻影响着统计、自动模式识别、数据挖掘等众多学科。当前相似性度量模型归为 4 类：几何途径相似性度量模型，基于特征的相似性度量模型，基于排列的相似性度量模型，基于变换的相似性度量模型。下面给出形式化的相似度量模型——正则相似度量模型(Regular Similarity Model, RSM)的形式化描述。

定义 4 正则相似模型记为 $RSM=<S,T,f_c,Sim>$，其中，$S=\{s_1,s_2,\cdots,s_n\}$ 为供应链产品符号序列集；$T=\{\tau_1,\tau_2,\cdots,\tau_m\}$ 为相似变换集；$f_c:T\times S\to[0,1]$ 为代价函数；$Sim(s_1,s_2)=\varphi(s_1,s_2)+\gamma(s_1,s_2)$ 为相似性度量，$\varphi(s_1,s_2)=1-C^{-\omega 1|LCS(s_1,s_2)|}$，称为同构相似性，$\gamma(s_1,s_2)=C^{-\omega 2\min(f_c(\tau 1)+f_c(\tau 2))}$，称为异构相似性，其中，$\tau_1,\tau_2\in T$，$C,\omega_1,\omega_1$ 为常数，$C\in(1,\infty)$，$\omega_1,\omega_1\geq 0$，且 $\tau_1(s_1)=\tau_2(s_2)$。RSM 的相似变换集与代价函数可根据具体问题指定。

定义 5 给定 Σ 上供应链产品符号序列集 S，其中，$s,t\in S$。供应链产品符号序列 s 和 t 的编辑距离可递归定义为

$$d_{edit}(s,t)=\min\begin{cases}cost(Del(t[1]))+d_{edit}(s,Rest(t))\\cost(Del(s[1]))+d_{edit}(Rest(t),t)\\cost(Sub(s[1],t[1]))+d_{edit}(Rest(s),Rest(t))\end{cases}\quad(1)$$

其中，$Del(s[1])$ 表示删除 s 中第 1 个字符；$Sub(s[1],t[1])$ 表示替换 s 和 t 中第 1 个字符；$Rest(s)$是 s 去掉第 1 个字符的子串；$cost$ 是执行删除或替换操作的代价。

定义 6 供应链产品符号序列 s 到供应链产品符号序列集 S 的距离定义为

$$D_{edit}(s,S)=\sum_{s^*\in S}d_{edit}(s,s^*)\quad(2)$$

定义 7 供应链产品符号序列 s 到有限产品符号序列集 S 的最大距离：

$$\Delta_{dist}(s,S)=\max_{s^*\in s}d_{dist}(s,s^*)\quad(3)$$

对 $\forall s^*\in S,\ \exists s_c\in\Sigma^*$，满足 $\Delta_{dist}(s_c,S)\leq\Delta_{dist}(s^*,S)$，则 s_c 为 S 的中心。

因为求两序列编辑距离问题有多项式时间复杂度算法，所以求给定产品符号序列集的中位序列是多项式时间内可得的。简单的方法是取每一产品符号序列到产品符号序列集中其他序列间的编辑距离之和中最小的那个序列，其时间复杂度为 $O(n^2)$。然而，求中值序列的问题却没有多项式时间内的有效算法。

3 供应链产品符号序列的正则相似度量模型

在给出 RSM 模型定义后，需要解决以下问题：

(1)对长度有限的任意两产品符号序列，RSM 描述的正则相似性度量是否总能被定义，如果不能，那么需要什么条件。

(2)相似性变换和代价函数定义，对 RSM 模型输出两序列间相似性度量的值的大小有何影响。

为此，给出供应链产品符号序列正则相似度量模型性质。

定义 3 要求相似变换集中至少存在变换使产品符号序列集中任两产品符号序列能够转换为相同的产品符号序列。因此，对没有定义合适相似变换的 RSM，不能给出任意产品符号序列相似性度量定义。为确保 RSM 有效，给出如下定理：

定理 1 $s_1,s_2\in S,\ \exists T_m$，使 $\tau 1(s_1)=\tau_2(s_2)$，其中，$\tau_1,\tau_2\in T_m$。

证明：设 $T=\{\tau_{ins},\tau_{del},\tau_\phi\}$，$\tau_{ins},\tau_{del}$ 分别表示插入、删除一个字符的相似性变换，τ_ϕ 表示空变。T^*是 T 的闭包。现做如下相似变换，考虑依次删除 s_1 中所有字符，然后依次插入 s_2 中每个字符：$\tau:(\tau_{del}<.k_1.>^*.\tau_{ins}<s_2^{11}>.\tau_{ins}\cdots\tau_{ins}<s_2^{|s_2||s_2|}>)$，$\tau\in T^*$，因此，$\tau(s_1)=\tau_\phi(s_2)$。$T^*$就是寻找的 T。

由定理 1 证明，S 中任何产品符号序列都可在 T^*找到合适的相似性变换，使其变为 S 中的另一产品符号序列。

很难对 T^*中的每个相似性变换规定变换代价函数，但可以在 T 基础上定义变换代价函数 f_c。

设基本变换 $T=\{\tau_{ins},\tau_{del},\tau_\phi\}$，$\tau_{custom}$ 为可自定义的相似性变换，各相似变换代价函数定义见表 1。

表 1 相似变换代价函数

相似变换 $s_2=\tau_e(s_1)$	变化代价函数 f_c
τ_{ins}	1
τ_{del}	1
τ_ϕ	0
τ_{custom}	自定义
$\tau_e:e(\tau_{ins},\tau_{del},\tau_\phi)$	$\sum_{k\in S}f_c(\tau_{ins}<k>,S)+\sum_{k\in S}f_c(\tau_{del}<k>,S)$

4 供应链产品符号序列相似性算法

两产品符号序列的 RSM 相似度量相似性计算问题实际是求 RSM 同构相似性 $\varphi(s_1,s_2)=1-C^{-\omega 1|LCS(s_1,s_2)|}$ 和异构相似性 $\gamma(s_1,s_2)=C^{-\omega 2\min(f_c(\tau_1)+f_c(\tau_2))}$。由于其与最大公共子序列问题本质的类似，因此可以用动态规划的办法求解。设供应链产品符号序列 $s<s_1,s_2,\cdots,s_m>,t<t_1,t_2,\cdots,t_n>$ 的最大公共子序列 $LCS(s,t)$记为 $z=<z_1,z_2,\cdots,z_k>$。

如果 $s_m=t_n$，则 $z_k=s_m=t_n$，且 $z^{(k-1)}$ 是 $s^{(m-1)}$ 和 $t^{(n-1)}$ 的最大公因子序列，其中，$z^{(k-1)}=<z_1,z_2,\cdots,z_{k-1}>$；$s^{(m-1)}=<s_1,s_2,\cdots,s_{m-1}>$；$t^{(n-1)}=<t_1,t_2,\cdots,t_{n-1}>$。

如果 $s_m\neq t_n$，且 $z_k=s_m=t_n$，则 z_k 是 $s^{(m-1)}$ 和 t 的最大公因子序列；如果 $z_k\neq t_n$，则 z_k 是 $t^{(n-1)}$ 和 s 的最大公因子序列。

有了上面的递归规律，可以设计算法计算出 2 个产品符号序列的最大公共子序列，并进而得出从 s 到 t 的相似变换序列。为此，给出正则相似度量模型 RSM 相似性算法如下：

输入 产品符号序列 $s<s_1,s_2,\cdots,s_m>,t<t_1,t_2,\cdots,t_n>$

输出 $LCS(s,t)$和 s 到 t 的相似性变换序列 τ

Step1 设 c[i,j]用来保存 $s^i<s_1,s_2,\cdots,s_i>,t^j<t_1,t_2,\cdots,t_j>$ 的最大公共序列的长度 $|LCS(s^i,t^j)|$。

Step2 如果 $i>m$，则 goto Step 5。

Step3 如果 $s_i=t_j$，则 $c[i,j]=c[i-1,j-1],b[i,j]=NW$，否则，如果 $c[i-1,j]\geq c[i,j-1]$，$c[i,j]=c[i-1,j],b[i,j]=N$，则 $c[i,j]=c[i,j-1],b[i,j]=W;\ j\leftarrow j+1;i\leftarrow i+1$。

Step4 根据记录回溯，反求出 LCS(s,t)到 t 的相似性变换序列，即调 LCSBacktrace (b,lcs,m,n)输出 lcs。

Step5 调用 TB (b,transforms,m,n)，输出 transforms。

下面分别给出 LCSBacktrace 和 TB 的子过程：

(1)LCSBacktrace 的子过程

```
LCSBacktrace(b,X,i,j)
Step1 if i=0 或 j=0,then 返回；
Step2 if b[i,j]=NW,then LCSBacktrace(b,X,i-1,j-1),  X ← X+s_t
        else if b[i,j]=N,then LCSBacktrace(b,X,i-1,j)
            else LCSBacktrace(b,X,i,j-1)
```

(2)TB 子过程

```
TBacktrace(b,i,j,s,t)
Step1 if i=0 且 j=0,then 返回；
Step2 if b[i,j]=NW,then TB (b,i-1,j,s,t);  X ← X + τ_del(S_i)
        else TB (b,i,j-1,s,t); X ← X + τ_ins(t_j)
```

针对供应链企业，经常需要处理一类重要的数据——偏好序列数据。数据整理后可得到偏好产品符号序列数据：被评价对象(产品)组成的有限符号表 $\Sigma=\{A_1,A_2,\cdots,A_m\}$，被调查者(评价主体)组成的产品符号序列集 $S=\{s_1,s_2,\cdots,s_n\}$，S 中每个序列对应被调查者对 Σ 中产品打分的一个排序。

假设给定所示偏好产品符号序列集，考虑根据被调查者的偏好序列来描述被调查者 $P_1,P_2,\cdots,P_5$ 之间的相似性，可能需要回答<A_1,A_2,A_3>和<A_2,A_3,A_3>相似程度与<A_1,A_2,A_3,A_4>和<A_2,A_3,A_3,A_4>相似程度是否一样。该应用的数据分析人员有理由认为<A_2,A_3,A_3,A_4>自相似程度比<A_1,A_2,A_3>大一些。因此，可以设置 RSM 的参数 c 为 1.1 和 1.2 等较小的值。按 RSM 相似性度量定义，得出各序列间相似性度量的值，与直觉吻合得很好，如表 2 所示。

表 2　偏好序列集 S 任意两序列的 RSM 度量

偏好序列	RSM 相似度量(c=1.2)				
	s_1	s_2	s_3	s_4	s_5
s_1=<$A_1A_2A_3$>	1.413				
s_2=<$A_2A_1A_3$>	1.000	1.413			
s_3=<$A_3A_2A_1$>	0.652	1.000	1.413		
s_4=<$A_1A_2A_3A_4$>	1.261	0.874	0.565	1.518	
s_5=<$A_4A_3A_2A_1$>	0.565	0.874	1.261	0.510	1.518

相比之下，如果使用产品符号序列常用的编辑距离或海明距离，都不能很好吻合这个具体应用对相似性的直觉认识。

5　供应链产品偏好市场可视化

通过 RSM 相似度量模型，能够得出神经元之间的相似性，根据每个神经元对应产品符号序列。把神经元和其特征产品符号序列中每个符号(对应不同产品)都显示在二维图表上，图中的圆圈对应 SOM 退火符号模型的单个神经元，圆圈大小对应神经元被命中次数，图中用方块表示不同的产品。通过圆圈与方块间位置关系表示消费群体对某产品的偏好程度和差异性等，如图 1 所示。

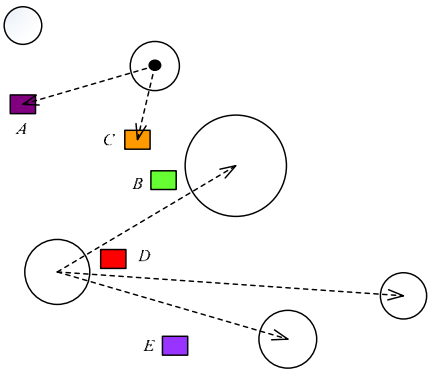

图 1　产品偏好市场分布(产品需求)

6　结束语

本文借鉴数据挖掘的产品符号序列聚类方法，提出了产品符号序列聚类的相似性度量模型(RSM)，进一步提出基于 RSM 模型的模拟退火算法，通过调整 RSM 模型参数，RSM 可以变为与编辑距离、海明距离等价的相似性度量，将该模型应用于 SOM，该算法可有效解决随消费者市场偏好变化的产品最佳配置问题，给出利用偏好数据进行市场细分和产品配置调整的较为完整的技术实现途径，科学部署与实施新产品结构和销售网点，从而确定有效的供应链分销网络。

参考文献

[1] Hsu Tsuen-Ho. The Fuzzy Clustering on Market Segment[C]//Proc. of the 9th IEEE International Conference. [S. 1.]: IEEE Press, 2005: 621-626.

[2] Hruschka H. Comparing Performance of Feedforward Neural Nets and K-means for Cluster-based Market Segmentation[J]. European Journal of Operational Research, 2004, 114(2): 346-353.

[3] Kuo R J. Integration of Self-organizing Feature Map and K-means Algorithm for Market Segmentation[J]. Computers & Operations Research, 2004, 29(1): 1474-1493.

[4] LeBlanc L J, Galbreth M R. Designing Large-scale Supply Chain Linear Programs in Spreadsheets[J]. Communications of the ACM, 2007, 50(8): 59-64.

[5] Neal W D. Advances in Market Segmentation[J]. Marketing Research, 2001, 13(1): 14-18.

编辑　索书志