第 35 卷 第 1 期 计 算 机 工 程 2009 年 1 月
Vol.35 No.1 Computer Engineering January 2009
·人工智能及识别技术· 文章编号：1000—3428(2009)01—0178—02 文献标识码：A 中图分类号：TP391

基于数据挖掘的符号序列聚类相似度量模型

郑宏珍，初佃辉，战德臣，徐晓飞
(哈尔滨工业大学智能计算中心，264209)

摘 要：为了从消费者偏好序列中发现市场细分结构，采用数据挖掘领域中的符号序列聚类方法，提出一种符号序列聚类的研究方法和框架，给出 RSM 相似性度量模型。调整 RSM 模型参数，使得 RSM 可以变为与编辑距离、海明距离等价的相似性度量。通过 RSM 与其他序列相似性度量的比较，表明 RSM 具有更强的表达相似性概念的能力。由于 RSM 能够表达不同的相似性概念，从而使之能适用于不同的应用环境，并在其基础上提出自组织特征映射退火符号聚类模型，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：符号序列聚类；数据挖掘；相似性模型

Symbolic Sequence Clustering Regular Similarity Model Based on Data Mining

ZHENG Hong-zhen, CHU Dian-hui, ZHAN De-chen, XU Xiao-fei
(Intelligent Computing Center, Harbin Institute of Technology, Harbin 264209)

【Abstract】From a consumer point of the sequence of preference, data mining is used in the field of symbolic sequence clustering methods to detect market segmentation structure. This paper proposes a symbolic sequence clustering methodology and framework, gives the similarity metric RSM model. By adjusting RSM model, parameters can be changed into RSM and edit distance, Hamming distance equivalent to the similarity metric. RSM is compared with other sequence similarity metric, and is more similar to the expression of the concept of capacity. As to express different similarity, the concept of RSM can be applied to different applications environment. Based on the SOM annealing symbol clustering model, the consumer preference for market segmentation can be studied in the structure, which means it is realized in practical application.

【Key words】 symbolic sequence clustering; data mining; similarity model

1 概述

在经济全球化的环境下，面对瞬息万变的市场和技术发展，企业要想在国内外市场竞争中立于不败之地，必须对客户和市场需求做出快速响应。目前，通过市场调研公司或企业自身的信息系统，收集来自市场和消费者的数据相对容易，而如何理解数据反映的市场细分结构和需求规律却是相当困难的。

为解决这一问题，许多研究者选择消费者的职业、收入、年龄、性别等特征数据作为细分变量，利用统计学传统聚类方法得到市场细分结构[1-2]。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]。

为此，本文从用户偏好序列数据对市场进行细分。通过对符号序列数据相似性的研究，给出一个可形式化的 RSM 相似性度量模型和算法概要。该度量模型考虑了 2 对象之间相似与相异 2 个方面的因素，通过参数的调整，可以根据问题的具体性质表达不同的相似性概念。并在此基础上，将在数值型数据领域表现良好的 SOM 神经网络引入到符号序列数据的聚类问题上，给特征符号序列的机器自动识别提供了可能性。

2 符号序列聚类问题

序列聚类问题作为发现知识的一种重要的探索性技术，受到数据挖掘与知识发现研究领域的极大重视。企业决策者在进行市场和产品相关战略时，迫切需要某些技术手段来理解序列数据，这也正是本文研究的序列聚类问题的工程背景。

下面给出符号序列的相关定义。

定义 1 设 $A=\{a_1,a_2,\cdots,a_n\}$ 为有限符号表，A 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 $|s|$。A 上所有有限长度符号序列集合记为 A^*。例如：符号表$\{a, b, c, d, e, f, g\}$，则$<abf>$, $<cdbg>$是符号序列。

定义 2 设 $P=\{S_1,S_2,\cdots,S_t,\cdots,S_n\}$，$S_t$是$A^*$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。

3 符号序列的正则相似度量模型

相似性度量往往与问题的应用背景具有紧密联系，并影响符号序列聚类结果。为此建立符号序列形式化的相似性度量模型，并在此基础上研究符号序列的聚类问题。

3.1 正则相似度量模型

下面给出形式化的相似度量模型——正则相似度量模型

基金项目：国家“863”计划基金资助项目“CIMS模型驱动的智能化软构件与软件生成技术”(2006AA01Z167)
作者简介：郑宏珍(1967－)，女，副教授，主研方向：数据挖掘，智能计算；初佃辉，副教授、硕士；战德臣、徐晓飞，教授、博士
收稿日期：2008-06-24 **E-mail：**hithongzhen@163.com

(Regular Similarity Mode, RSM)的形式化描述。

定义 3 给定相似变换集合 $T=\{\tau_1,\tau_2,\cdots,\tau_m\}$，符号序列集 $S=\{s_1,s_2,\cdots,s_n\}$，变换代价函数定义为 $f_c:T\times S\rightarrow [0,1]$。

定义 4 $s_1,s_2\in S,\tilde{S}_{12}$ 是s_1, s_2子序列集合，设 $\tilde{s}_z\in\tilde{S}_{12}$，若 $\forall\tilde{s}\in\tilde{S}_{12}$ 满足 $|\tilde{s}_z|\geq|\tilde{s}|$，称 $\tilde{s}_z$ 是s_1, s_2的最大公共子序列，记为 $LCS(s_1, s_2)$。

定义 5 正则相似模型记为 $RSM=<S,T,f_c,Sim>$。其中，符号序列集 $S=\{s_1,s_2,\cdots,s_n\}$；相似变换集 $T=\{\tau_1,\tau_2,\cdots,\tau_m\}$；代价函数 $f_c:T\times S\rightarrow[0,1]$；相似性度量 $Sim(s_1,s_2)=\alpha(s_1,s_2)+\beta(s_1,s_2)$。式中，$\alpha(s_1,s_2)=1-C^{-\delta_1|LCS(s_1,s_2)|}$ 称为同构相似性；$\beta(s_1,s_2)=C^{-\delta_2\min(f_c(t_1)+f_c(t_2))}$ 称为异构相似性，$\tau_1,\tau_2\in T$，δ_1,δ_2 为常数，$C\in(1,\infty)$，$\tau_1(s_1)=\tau_2(s_2)$。

RSM 的相似变换集与代价函数可根据具体问题而定。

3.2 正则相似度量模型性质

在给出 RSM 模型定义后，需对长度有限的任意两符号序列进行有效性分析，并根据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影响进行分析。下面给出 RSM 有效性前提。具体如下：

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

(1)对称性。

满足有效性前提的 RSM，$\tau_1\in\{\tau'\mid f_c(\tau')\}=\min_{\tau(s_1)=s_2}(f_c(\tau))$，$\tau_2\in\{\tau'\mid f_c(\tau')\}=\min_{\tau(s_2)=s_1}(f_c(\tau))$，则 $f_c(\tau_1)=f_c(\tau_2)$。

(2)传递性。

满足有效性前提的 RSM，对于 $\forall s_1,s_2\in S$，$\tau\in\exists T'$，使 $f_c(\tau)=\min(f_c(\tau_1)+f_c(\tau_2))$。

(3)排列不变性。

对于任意符号序列s_1, s_2，按任意方式对齐，式 $2n+2m+k=|s_1|+|s_2|$ 成立，n是s_1, s_2中不匹配符号的数量，m是匹配符号的数量，k是未找到对应符号的数量。

3.3 符号序列相似性变换

两符号序列的 RSM 相似度量相似性计算问题实际是求 RSM 同构相似性 $\alpha(s_1,s_2)$ 和异构相似性 $\beta(s_1,s_2)$。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。

设符号序列 $s<s_1s_2\cdots s_m>$，$t<t_1t_2\cdots t_n>$ 的最大公共子序列 $LCS(s, t)$记为 $z<z_1z_2\cdots z_k>$。并且：

$$z^{k-1}=z<z_1z_2\cdots z_{k-1}>,\ s^{m-1}=s<s_1s_2\cdots s_{m-1}>,$$
$$t^{n-1}=t<t_1t_2\cdots t_n>$$

如果 $s_m=t_n$，则 $s_m=t_n=z_k$，并且 z^{k-1} 是 s^{m-1} 和 t^{n-1} 的最大公共子序列。否则：

如果 $z_k\neq s_m$，则 z_k 是 s^{m-1} 最大公共子序列；

如果 $z_k\neq t_n$，则 z_k 是 t^{n-1} 和 s 的最大公共子序列。

根据以上递归规律，可以设计算法计算出 2 个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。

3.4 RSM 与其他序列相似性度量的比较

海明距离和编辑距离是目前较常用的符号序列相似性度量[4]。对于两等长符号序列，它们之间有越多的对应位置符号不同，则海明距离越大。编辑距离是将一个符号序列经插入、删除、替换等编辑操作变为另一个序列所需的操作次数。RSM所描述的符号序列的相似性由 α和β 两方面因素共同决定。$\alpha(s_1,s_2)$ 描述的是两符号序列间的共性，由两符号序列的最大公共子序列的长度的增函数表征；$\beta(s_1,s_2)$ 描述的是两符号序列间的差异，由符号序列间进行相似变换的最小代价的减函数表征。

设置 RSM 的参数 δ_1,δ_2，RSM 的相似性度量将与海明距离和编辑距离相对应。$\delta_1=0$ 时，有

$$Sim(s_1,s_2)=\alpha(s_1,s_2)+\beta(s_1,s_2)=C^{-\delta_2\min(f_c(\tau_1)+f_c(\tau_2))} \quad (1)$$

式(1)表明由 RSM 定义的相似性度量，如果定义的 $\tau1,\tau2$ 和对应的代价函数 f_c 与编辑距离定义的编辑变换一致，则 RSM 相似性度量与编辑距离的定义所表达的概念是一致的。$\delta_2=0$ 时，有

$$Sim(s_1,s_2)=\alpha(s_1,s_2)+\beta(s_1,s_2)=2-C^{-\delta_1|LCS(s_1,s_2)|} \quad (2)$$

式(2)表明由 RSM 定义的相似性度量表达的含义是 2 个符号序列对应位置相同的符号越多，则 2 个符号序列越相似，这与海明距离定义所表达的概念是一致的。

由于 RSM 中对相似性度量的定义包含了同构和异构 2 个部份，而海明距离或者编辑距离只描述了相似性中异构的那一部份，在实际应用情况下，被研究对象的相似性概念往往会同时由 2 部份概念组成，即它们之间有多少部份是共同的，它们之间还有多少部份是差异的，2 部份形成对不同对象之间相似性的印象，因此 RSM 具有更强的表达相似性概念的能力。

4 SOM 符号序列聚类算法

算法步骤如下：

Step1 构造 SOM 网络 *neurons*[*row*, *column*]，随机选择 S 中符号序列，作为各输出层神经元特征符号序列：

$neurons[i, j]\leftarrow random\_select(S)$

Step2 计算输出层神经元欧式距离矩阵 ***Ud***；

Step3 调用 TrainSOM 子过程训练网络；

Step4 返回训练后的 SOM 网络 *neurons*。

TrainSOM 子过程构造步骤如下：

TrainSOM(*sample*, *neurons*, ***Ud***, *step*)

Step1 计算 *sample* 与 *neurons* 中所有序列的 RSM 相似性度量，选出最相似于 *sample* 的神经元作为最匹配神经元：

$$BMU\leftarrow\max_{s_i\in neurons}(RSM(sample,s_i))$$

Step2 根据神经元距离欧式矩阵 ***Ud***，计算每个神经元分配的能量：

$$h(neuron_c)\leftarrow E_c\cdot e^{\frac{-\|neuron_c-neuron_b\|^2}{2\sigma^2}}$$

Step3 训练样本进行若干次相似变换，原特征符号序列被相似变换得到的新符号序列替换。

5 结束语

本文提出了形式化的符号序列相似性模型 RSM。该模型定义的符号序列相似性是由 2 个符号序列共同部份与差异部份两方面组成的度量值，能够表达不同的相似性概念，适用于不同的应用环境。在此基础上，研究对符号序列数据进行聚类的算法，提出 SOM 退火符号模型。该聚类算法通过调整输出层神经元网格数目，能够满足聚类粒度可以调整的要求。实验结果说明，SOM 退火符号模型聚类结果是比较稳定的，无论数据集合顺序如何，结果总趋向一致。

（下转第 194 页）