基于隐马尔可夫模型的符号序列自组织聚类

吕 昱 程代杰
(重庆大学计算机学院 重庆 400044)

摘 要 本文提出一种基于模型的、适合变长符号序列的自组织聚类算法。隐马尔可夫模型被用于表达各个聚类，批处理自组织特征被用于符号序列的聚类过程。实验结果表明该算法能有效发现变长符号序列中的聚类模式。

关键词 批处理自组织特征映射，隐马尔可夫模型，符号序列聚类

HMM Based Symbolic Sequence Self Organizing Clustering

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

Abstract In this paper, we propose a model-based, self organizing feature map algorithm for the clustering of variable-length sequences. Hidden Markov models(HMMs) are used as representations for the cluster centers, and batch map training algorithm is applied in clustering procedure. Simulation results show that our method can successfully find patterns of clusters of the input variable-length sequences.

Keywords Batch map, SOM, Hidden markov model, Symbolic sequence clustering

1 引言

聚类是人们认识未知事物，建立知识体系的重要方法。符号序列是人类生产、生活及科研工作中常见的一类数据。符号序列聚类问题随着近年来市场营销、Web 文本挖掘、分子生物学等领域迅速发展，成为数据挖掘领域一个研究热点。从方法论角度，符号序列聚类研究途径可以简单归为两大类：第一类研究途径是根据序列对象特点，定义相似性度量，然后应用通常的聚类算法对符号序列进行聚类[1,2]。这类途径往往需要依赖所研究问题的相似性的具体概念设计特别的聚类算法，而且多限用于固定长度符号序列的聚类。另一类研究途径是基于模型的聚类框架。这一研究途径用不同的概率模型(或其它模型)来表征不同聚类。而对符号序列来说，隐马尔可夫模型(HMM)是较为著名的一类模型[3,4]。隐马尔可夫模型能较好表达不同长度的符号序列，而且在语音、文本识别等实际问题中取得较成功应用。

文[5～7]将隐马尔可夫模型用于符号序列的聚类上，其基本思路是认定所研究数据来自于包含 K 个不同概率分布总体的混合模型，满足独立同分布条件。虽然 K 的大小和这 K 个组的概率分布及参数对于研究者是未知的，但是通过已知的数据和统计学习方法，比如期望最大算法，获得模型的参数。

隐马尔可夫模型聚类存在几个需要解决的难题：聚类的数目 K 如何确定；隐马尔可夫模型表达的聚类是不透明的；模型与模型之间的差异不容易表达。

对于聚类数目不易确定的问题，文[8]的解决办法是尝试不同数目 HMM，在得到的结果中，找到其中使得衡量聚类间差异的某特定指数最大的，作为最佳聚类方案。

对于聚类结果不透明这一问题，由于隐马尔可夫模型实际是描述一系列隐状态转移概率和隐状态激发可见符号概率，模型实质是序列的概率分布，对隐马尔可夫模型之间的相似性，文[9]做了初步的研究。

本论文提出基于隐马尔可夫模型的符号序列自组织聚类，即利用隐马尔可夫模型表达聚类，利用批处理自组织特征映射聚类算法对符号序列进行聚类，避免聚类数目 K 不易确定的问题，并引入隐马尔可夫模型间相似性度量的办法，从而可观察聚类的最终结果。本文第 2 节和第 3 节分别介绍了批处理自组织特征映射和隐马尔可夫模型；第 4 节提出基于隐马尔可夫模型的符号序列自组织聚类；第 5 节是对本文提出的符号序列聚类的仿真实验；最后是全文的结论。

2 批处理自组织特征映射

批处理自组织特征映射是 Kohonen 提出的一种神经网络训练算法。Kohonen 将该算法用于符号序列的聚类。算法大致步骤如下：

1)在训练样本中选择符号序列作为网络中神经元的特征符号序列；

2)在样本中寻找每个神经元 u_i 的特征符号序列的最近邻序列，形成符号序列集合 c_i；

3)对每个神经元 u_i，求其领域 N_i 神经元 u_j 对应的符号序列集合 c_j 所有的符号序列的中值序列，并以之作为神经元 u_i 的新的特征符号序列；

4)从步骤 2 重复执行一定次数，直到神经元的特征符号序列不发生变化。

Kohonen 尝试将该算法对大量独立单词进行聚类和分类识别，以用于语音识别中。

3 隐马尔可夫模型

离散隐马尔可夫模型包括下面几个部分：

1)模型的 N 种隐状态 $\{S_1, S_2, \cdots, S_N\}$，其中在时刻 t 所

吕 昱 博士生，主要研究方向为符号序列聚类技术。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。

处状态记为 q_t；

2)在各状态下，模型可能对外输出的可见符号组成的 M 个离散符号集$\{v_1, v_2, \cdots, v_M\}$；

3)模型各隐状态间转移概率分布 $A=\{a_{ij} \mid a_{ij}=P(q_{t+1}=S_j \mid q_t=S_i), i\leqslant i,j\leqslant N\}$；

4)模型在各隐状态输出不同可见符号的概率分布

$B=\{b_j(k) \mid b_j(k)=P(v_k at\ t \mid q_t=S_j) 1\leqslant j\leqslant N, 1\leqslant k\leqslant M\}$；

5)初始状态概率分布 $\pi=\{\pi_i \mid \pi_i=P(q_1=S_i), 1\leqslant i\leqslant N\}$。

给定隐马尔可夫模型 $\lambda=(A,B,\pi)$，其产生的符号序列记为 $O=O_1O_2\cdots O_T, O_t\in V$，定义 $\xi t(i,j)=P(q_t=S_i, q_{t+1}=S_j \mid O,\lambda)$，$\gamma_t(i)=\sum_{j=1}^{N}\xi t(i,j)$，则给定序列集$\{O^{(i)}\}_{i=1}^{k}$，对 a_{ij}，b_j 的估计可以通过 Baum-Welch 方法[3]得到：

$$\tilde{a}_{ij}^{k}=\frac{\sum_{l=1}^{k}\sum_{t=1}^{T^{(l)}-1}\xi_t(i,j \mid O^{(l)},\lambda)}{\sum_{l=1}^{k}\sum_{t=1}^{T^{(l)}-1}\gamma_t(i \mid O^{(l)},\lambda)} \tag{1}$$

$$\tilde{b}_j^{k}(v)=\frac{\sum_{l=1}^{k}\sum_{t=1\&O_t^{(l)}=v}^{T^{(l)}}\gamma_t(j \mid O^{(l)},\lambda)}{\sum_{l=1}^{k}\sum_{t=1}^{T^{(l)}}\gamma_t(j \mid O^{(l)},\lambda)} \tag{2}$$

4 HMM-SOM 混合聚类

C. de la Higuera 等的研究证明[11]，给定符号序列集$\{O^{(i)}\}_{i=1}^{k}$，寻找该序列集“中值序列”是 NP 难的。因此，Kohonen 提出的批处理自组织特征映射聚类算法在应用于符号序列聚类时，算法第 3 步对每个神经元 u_i，求其领域 N_i 神经元 u_j 对应的符号序列集合 c_j 所有的符号序列的中值序列，并以之作为神经元 u_i 的新的特征符号序列，是不存在多项式时间解的，对这一问题，Kohonen 及后续研究者只是启发性地给出求中值序列的办法[10~12]。本文为解决这一问题，提出以隐马尔可夫模型表示每个神经元的 HMM-SOM 混合聚类算法。算法基本框架如下。

1)根据 SOM 输出层神经元数量，假设为 c 个神经元，初始化隐马尔可夫模型 $\lambda_1, \lambda_2, \cdots \lambda_c$；

2)按 $d(O^{(k)} \mid \lambda)=-\log P(O^{(k)} \mid \lambda)$，计算每个训练样本最近邻的神经元 λ_i，形成符号序列集合 c_i；

3)对每个神经元 λ_i，求其领域 N_i 神经元 λ_j 对应的符号序列集合 c_j 所有的符号序列，以这些符号序列重新训练 λ_i；

4)从步骤 2 重复执行一定次数，到系统收敛或满足设定的终止条件。

HMM-SOM 混合聚类算法结束后，各神经元对应的隐马尔可夫模型 λ_i 相互间的差异程度需要能够引入合适的度量，以便于观察存在的聚类。

设 $\lambda_1=(A_1,B_1,\pi_1), \lambda_2=(A_2,B_2,\pi_2)$其中，

$$A_1=\begin{bmatrix} p & 1-p \\ 1-p & p \end{bmatrix}, B_1=\begin{bmatrix} q & 1-q \\ 1-q & q \end{bmatrix}, \pi_1=[0.5\ 0.5]$$

$$A_2=\begin{bmatrix} r & 1-r \\ 1-r & r \end{bmatrix}, B_2=\begin{bmatrix} s & 1-s \\ 1-s & s \end{bmatrix}, \pi_2=[0.5\ 0.5]$$

则两个隐马尔可夫模型是否相似可以由它们产生符号序列的统计性质来衡量，比如产生符号序列 O_t 的期望来表达，即：对所有符号 v_k，$E(O_t=v_k \mid \lambda_1)=E(O_t=v_k \mid \lambda_2)$，由该式可得：

$$s=\frac{p+q-2pq-r}{1-2r}。$$

任取 $p=0.3, q=0.3, r=0.1$，可得 $s=0.4$。可见两个隐马尔可夫模型参数很不一样，却可能产生相似的符号序列，具有类似的统计性质。

定义 λ_1 到 λ_2 的距离为 $D(\lambda_1,\lambda_2)=\frac{1}{T}[\log P(O^{(2)} \mid \lambda_1)-\log P(O^{(2)} \mid \lambda_2)]$，其中 $O^{(2)}=O_1O_2\cdots O_T$。由于 $D(\lambda_1,\lambda_2)\neq D(\lambda_2,\lambda_1)$，因此可引入 λ_1 与 λ_2 的对称的相异性度量定义：

$$D_s(\lambda_1,\lambda_2)=\frac{D(\lambda_1,\lambda_2)+D(\lambda_2,\lambda_1)}{2} \tag{3}$$

对 HMM-SOM 混合聚类算法结束后得到的隐马尔可夫模型 $\lambda_1, \lambda_2, \cdots, \lambda_c$，给定符号序列 O，按式(3)可计算得出各隐马尔可夫模型间相异性度量矩阵，从而利用 Umatrix 或 Sammon 映射等可视化技术，得到聚类结果的图形。

HMM-SOM 混合聚类算法与 Kohonen 提出算法本质区别是其使用概率模型表达聚类而后者使用具体的中值序列表达聚类。HMM-SOM 的主要优点是避免求中值序列，算法花费时间主要是第 2、3 步，这两个步骤分别对应隐马尔可夫模型的估值问题和学习问题均具备多项式时间的求解算法，因此 HMM-SOM 算法时间复杂度是多项式时间可解的。

5 实验分析

实验数据是利用程序产生的人工数据集。共生成实验数据集 A 和 B 两组实验数据。

第一组实验数据生成方法是：设定 3 个不同符号序列，以这 3 个符号序列为基础，我们称其为核，对核正态随机地进行变换[1]，产生 3 个符号序列集，每个集合产生 100 个序列，然后混合这 3 个符号序列集的符号序列，得到实验数据集 A。

第二组实验数据生成方法是预设 3 个不同的隐马尔可夫模型，由 3 个不同的模型各自生成 100 个符号序列数据，混合得到实验数据集 B，参考图 1。

图 2 和图 3 是 HMM-SOM 混合聚类与直接批处理自组织特征映射聚类结果[2] 的对比。从实验结果可见，HMM-SOM 更好地发现了人工数据集合中的聚类模式。

结论 本论文应用自组织特征映射网络对变长符号序列聚类，与传统计算符号序列集的中值序列的启发算法不同，本文是通过引入隐马尔可夫模型表达聚类中心的，该算法具有多项式时间的计算复杂度，较传统基于中值序列的启发算法，具有更好实验效果。

参 考 文 献

1 Manning A, Brass C, Goble, Keane J. Clustering techniques in biological sequence analysis. In First European Symposium on Principles of Data Mining und Knowledge Discovery, 1997. 315~322

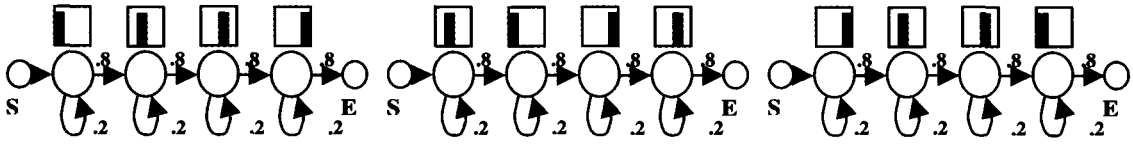
2 Vijaya P A, Murty M N, Subramanian D K. An Efficient Technique for Protein Sequence Clustering and Classification. In: Proc. of the 17th Intl. Conf. on Pattern Recognition, 2004

3 Rabiner L. A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE, 1989,

[2]“正态随机的进行变换”是指可重复地产生一组随机整数集$\{\cdots,-3,-2,-1,0,1,2,3,\cdots\}$，不同值的整数个数服从正态分布。对核变换的过程是取随机整数集中某个整数 i，对核随机删除或添加 i 个符号。比如当前取到的随机整数 -3，则随机删除核中的 3 个符号，从而得到一个符号序列样本，如果当前随机整数是 2，则在核中随机选择两个位置，插入符号表中两个符号；如果随机整数是 0，则不进行任何操作。

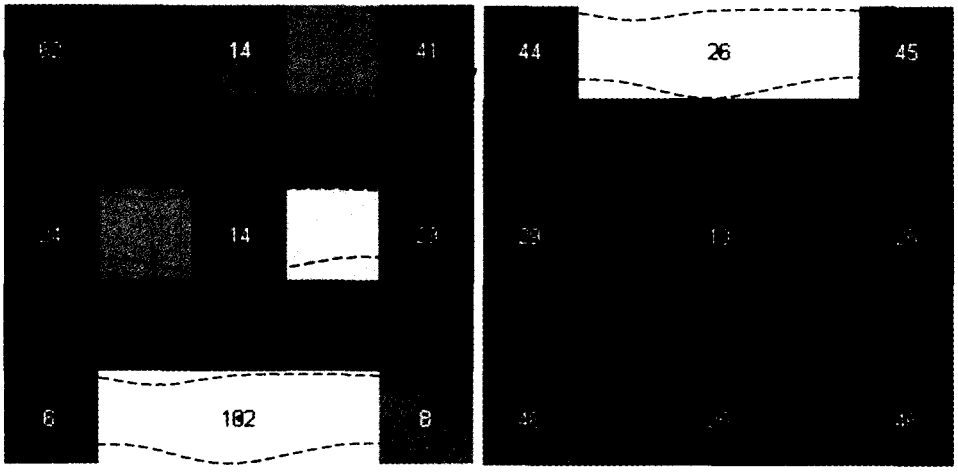
[3] 为了可视化批处理自组织特征映射聚类结果，使用神经元的特征符号序列间的编辑距离求得相邻神经元之间的距离矩阵。

77(2):257～286
4 Krogh A. An Introduction to Hidden Markov Models for Biological Sequences. In Computational Methods in Molecular Biology. Elsevier,1998. 45～63
5 Smyth P. Clustering sequences with hidden Markov models. Advances in Neural Information Processing Systems 9, 1997. 648～654
6 Owsley L, Atlas L, Bemard G. Self-organizing feature maps and hidden Markov models for machine-tool monitoring. IEEE Transactions on Signal Processing,1997
7 Cadez I, Heckerman D. Model-Based Clustering and Visualization of Navigation Patterns on a Web Site. Data Mining and Knowledge Discovery,2003,7:399～424
8 Li C, Biswas G. Temporal pattern generation using hidden Markov model based unsupervised classification. In: Proc. of the Third International Symposium on Intelligent Data Analysis,1999
9 Juang B H, Rabiner L R. A probabilistic distance measure for hidden Markov models. AT&T Tech. J,1985,64(2):391～408
10 Kohonen T. self-organizing maps. 3rd edition, Springer-Verlag, 2001. 206～207
11 de la Higuera C, Casacuberta F. Topology of strings: median string is NP-complete. Theoretical Computer Science,2000
12 Fischer I, Zell A. Processing Symbolic Data With Self-Organizing Mps,2000



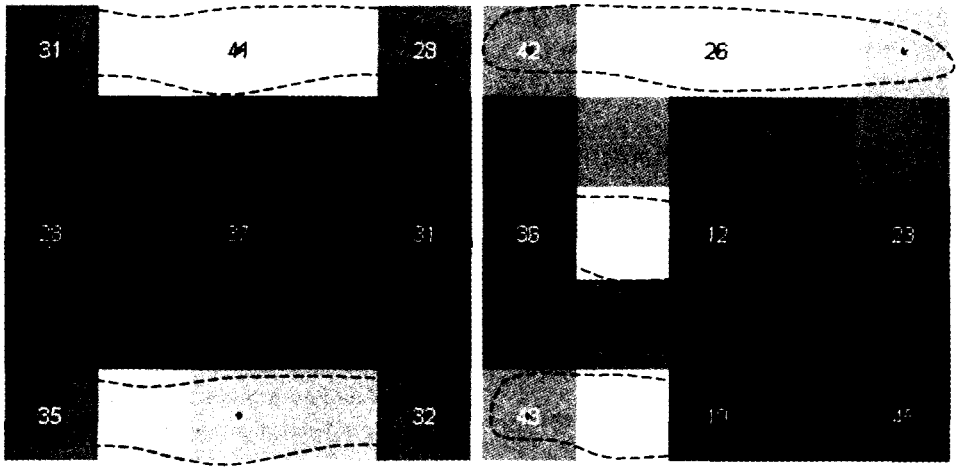
a) HMM1　　b) HMM2　　c) HMM3

图 1　产生人工数据集的 3 个隐马尔可夫模型



a) HMM-SOM 混合聚类　　b) 以编辑距离为相似度量的 SOM 符号聚类

图 2　实验数据集 A 聚类结果



a) HMM-SOM 混合聚类　　b) 以编辑距离为相似度量的 SOM 符号聚类

图 3　实验数据集 B 聚类结果