第35卷 第22期 计 算 机 工 程 2009年11月
Vol.35 No.22 Computer Engineering November 2009
·博士论文· 文章编号：1000—3428(2009)22—0026—03 文献标识码：A 中图分类号：TP391

基于最大熵投票模型的时间序列无监督分割

孙 焘，冯 林，郑 虎，高成锴
(大连理工大学创新实验学院，大连 116024)

摘 要：通过高维时间序列分割可以创建高级符号表示。提出一种针对高维时间序列的无监督分割算法，用于解决高维数据符号化的预处理问题。该算法实现对高维数据的聚类，应用最大熵投票模型进行序列分割。实验结果表明，其平均查全率和查准率分别为0.86和0.88，且整体性能优于主成分分析算法和概率主成分分析算法。

关键词：最大熵投票模型；*k*-mean聚类；高维时间序列；无监督分割

Unsupervised Segmentation of Time Series Based on Max Entropy Voting Model

SUN Tao, FENG Lin, ZHENG Hu, GAO Cheng-kai
(School of Innovation Experiment, Dalian University of Technology, Dalian 116024)

【Abstract】 Through the high-dimension segmentation, the high-level symbol expression can be created. This paper proposes an unsupervised segmentation algorithm for high-dimension time series. This method can solve the pretreatment problem of high-dimension symbolization. It realizes the clustering of high-dimension data, and uses max entropy voting model to do series segmentation. Experimental results show that the algorithm's average recall ratio and precision ration are respectively 0.86 and 0.88. Its whole performance is better than Principal Component Analysis(PCA) algorithm and Probability Principal Component Analysis(PPCA) algorithm.

【Key words】 max entropy voting model; *k*-mean clustering; high-dimension time series; unsupervised segmentation

1 概述

时间序列是一种数据序列，其中的数据随时间变化而变化，它反映了某种属性值随时间变化的特征。近年来，出现了许多复杂庞大的高维时间序列数据库，导致计算复杂度激增，使多数能成功应用于一维时间序列的挖掘技术无法用于挖掘高维时间序列。

在高维数据分割领域中，文献[1]研究了如何找到具有最优解的分割，其后关于一维时间序列的研究逐渐成为热点，很多一维时间序列分割方法被提出。文献[2]提出一种使用最大熵模型的算法，能处理文本类的分割。文献[3]使用概率主成分分析(Probability Principal Component Analysis, PPCA)算法分割捕捉序列，并对不同算法的结果进行比较。

国内的相关研究起步较晚，关于高维时间序列分割的研究很少。文献[4]提出一种评估时间序列分割结果以及分割算法性能的评价指标，但该评价方式不适合高维时间序列。文献[5]提出一种时间序列的稳健最优分割方法，其基本思想是基于对线性模型的数据矩阵进行奇异值分解(Sigular Value Decomposition, SVD)，以自适应确定子序列最合适的多项式阶次，从而避免穷举寻优过程，极大提高了计算效率，但以准确性为代价，其计算结果误差较大。

本文提出一种基于最大熵的时间序列分割算法，先对高维数据使用聚类化简，然后利用最大熵投票模型进行分割。该算法易于实现，具有无监督特性，可以应用在其他领域。

2 分割算法

2.1 聚类算法

对高维数据进行聚类化简，本文使用 *k*-mean 聚类算法。

给定需划分的聚类个数为 k，先得到 k 段初始划分的集合，然后采用迭代重定位技术，通过将对象从一个簇移到另一个簇来改进划分的质量。

算法流程如下：

(1)假设要聚成 k 个类，人为决定 k 个类中心 $Z_1(1)$，$Z_2(1),\cdots,Z_k(1)$。

(2)在第 k 次叠代中，样本集 $\{Z\}$ 用如下方法进行分类：对所有 $i=1,2,\cdots,k,i\neq j$，若 $\|Z-Z_i(k)\|<\|Z-Z_i(k)\|$，则 $Z_j(k+1)=Z_j(k)$，$Z\in S_j(k)$。

(3)令由(2)得到的 $S_j(k)$ 的新类中心为 $Z_j(k+1)$，令 $J_j=\sum_{Z\in S_j(k)}\|Z-Z_j(k+1)\|^2$ 最小，且 $j=1,2,\cdots,K$，则

$$Z_j(k+1)=\frac{1}{N_j}\sum_{Z\in S_j(k)}Z$$

其中，N_j 为 $S_j(k)$ 中的样本数。

(4)对于所有 $j=1,2,\cdots,K$，若 $Z_j(k+1)=Z_j(k)$，则终止，否则转(2)。

设高维时间序列为 L，$l_i(x_i,y_i,z_i,\cdots,k_i)\in L$，$l_i$ 是时间序列 L 的某个点，x_i，y_i，$z_i,\cdots,k_i$ 是点对应维度的坐标。例如，对于二维空间的点来说，是基于 x, y 坐标的点。二维空间聚类的结果是各自独立的圆。设 d 为圆的半径，即聚类中心点

基金项目：国家自然科学基金资助项目(60773213)；辽宁省自然科学基金资助项目(20071092)

作者简介：孙 焘(1976－)，男，博士，主研方向：高维时间序列数据挖掘，智能图像处理，机器视觉，演化算法；冯 林，教授、博士、博士生导师；郑 虎、高成锴，硕士研究生

收稿日期：2009-05-17 **E-mail：**gaochengkai@gmail.com

到边缘点的距离，聚类中心为 $m(a,b)$，则某个独立的类的几何形式可以表示为

$l\ (x-a)^2+(y-b)^2=d^2$

若在三维空间的聚类中心点为 $m(a,b,c)$，则三维空间某个聚类表示为

$(x-a)^2+(y-b)^2+(z-c)^2=d^2$

依此类推，高维空间聚类的结果是各自独立的高维超级球体，若聚类中心为 $m(a,b,c,\cdots,k)$，则对应的某个聚类可以表示为

$(x-a)^2+(y-b)^2+\cdots+(m-k)^2=d^2$

对每个超级球体使用符号进行标记，由于标记的符号对聚类没有影响，因此结果不取决于符号的不同。程序使用数字进行标记，球体内的任意高维点聚类后都使用球体本身的标号表示。

2.2 基于最大熵的投票专家分割算法

研究任意一个时间序列分段的前提都是其具有相应的含义和特征，不同的时间序列分段具有的消息含义不同。分割即把具有不同含义的段落区分出来，最大熵分割算法实现片段具有的 2 个属性：

(1)随机一致的几率很小，子串出现的频率对应其成为片段的可能性。

(2)序列中的片段，其后缀字符是多样的，能通过观察一个序列的左右临近序列辨别出一个片段。

投票专家算法在时间序列上滑动一个长度为 n 的窗口，在窗口的每个位置对所处片段进行属性值计算。算法使用树结构，储存所有片段的属性值。

建立一个 n 键树，深度为 n+1。i+1 层的树节点表示长度为 i+1 的序列。如图 1 所示，以 *abcbcd* 串为例，产生深度为 3 的一个键树。

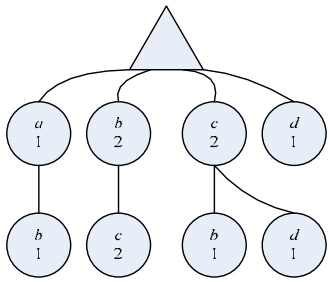

图 1　字符串 *abcbcd* 的键树

每个长度为 2 和长度小于 2 的序列被图 1 所示的树中的节点表示出来，其中，数字表示后继的发生次数，子串 *ab* 发生了一次，a 的后继节点 b 的键值为 1。对于一个任意变量 X，其离散的边界熵的计算公式为

$$H(X)=-\sum_{x\in X}p(x)\lg p(x)$$

通过键树能容易地计算边界熵。在图 1 中，节点 a 的边界熵是 0，因为它只有一个孩子，即 *ab*。节点 c 的边界熵是 1.0，因为它有 2 个等概率的孩子，即 *bc* 和 *bd*。树的根节点没有属性值，在深度为 n+1 的树中只有深度在 n 内的节点有熵属性值。所有子串的熵和频率值均经过标准化，用以公平地衡量偏离相同长度子串的距离。

在长度为 k 的序列中，存在 k-1 个判定位置作为边界的候选，算法使用长度为 n 的窗口，读入一定长度的字符，不同专家投票给窗口中具有相应最高属性值的判定位置，投票后窗口向后滑动，即最先读入的字符出列，并读入下一个字符。专家继续判定和投票，直到读入最后一个字符。

3　序列分割

每个在序列中潜在的分割位置获得一定数目的选票后，算法必须根据投票估计出边界，决定从哪里进行分割。这些票数在数学上相当于一列函数值，可以考虑取其极值，即如果一个潜在的边界有局部最大票数，则判定它为分割位置。算法也可以通过设定阈值，对一个边界的投票数目若超过该阈值则判定其为边界。

边界熵专家投票给边界熵最高的边界，频率专家尝试找到一个序列的最大概率的分割位置，当 2 位专家同时对一个边界投票，特别是对一个边界重复投票时，容易得到一个局部边界。此时，分割算法能很方便地确定哪一个是边界。

3.1 实验判定标准

在人体动作捕捉序列中，主要使用的衡量标准为查全率和查准率。

定义 a 为检索或通过程序得到的判定结果，b 为判定结果 a 中正确的、真正需要的部分，c 为判定或检索中的全集，d 为全集中所有正确的部分，则查全率为 b/d，查准率为 b/a。

在人体捕捉序列中，分割边界的定义为使用手工分割原始的动作序列，对跑跳走等情况一一列出。每个动作的停顿时间不一样导致动作分割边界无法被准确限定，将停顿时间长的边界被设定成阈值，以便对程序结果进行比较。图 2~图 4 给出了对具有 2 751 个时间点的跳跃序列的分割结果。

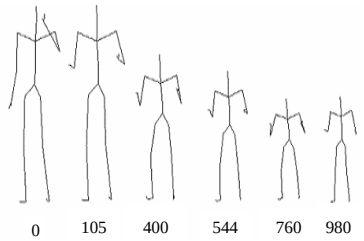

图 2　分割结果的背部视角

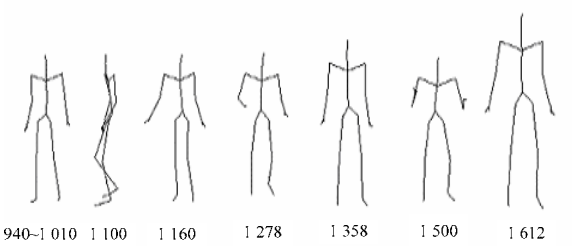

图 3　分割结果的正面视角

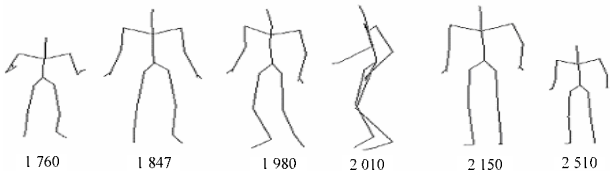

图 4　分割结果的侧面视角

可以看出，手工分割的目的是在重复动作之间和不同动作转接处进行划分。

3.2 聚类算法对分割结果的影响

聚类的数目对分割结果具有很大影响，表 1 描述了某个动作捕捉序列的聚类个数与聚类结果的关系，其中，源数据长度为 2 500，源数据已进行了手工分割；信息存留表示聚类个数与序列长度之比。通过实验得到聚类最优的数目为 50。

表 1　聚类个数与聚类结果的关系

聚类数	*a*	*b*	查准率	查全率	信息存留
10	12	17	0.7	0.6	0.004
50	11	22	0.5	0.6	0.020
100	17	20	0.8	0.9	0.040

聚类数目的不同决定了信息的损耗量，一个 2 500 时间长度的动作序列被 50 个序列表示，可以估计此序列的信息存留量为 1/50，若表示符号为 2 500 个，则信息存留量为 1。运动捕捉数据是高维的，在原始序列中几乎没有相同的点，仅存在距离很近的一些点，大量保存信息是不现实的，且没有很多能处理大量数据的方法。因此，信息存留量是相对不同聚类个数而言的。信息存留量值越高，相对保留的信息越多。

从算法角度来看，聚类个数的增加提高了时间序列中的不确定性，对应的熵和片段会增多，导致相应的最大熵值分布变得零乱，熵专家投票数变得分散，无法把选票全部投给正确边界。因此，通过实验得到聚类的最优数目为 50，信息存留量仅说明相对信息的存留量。

4　实验分割结果

实验采用长度为 30 的窗口，键树高度为 31，边界投票超过 15 票的点认定为分割点，序列为 10 组不同动作捕捉序列，均出自卡耐基梅隆大学的人体运动序列捕捉数据库。序列分割结果如表 2 所示。

表 2　序列分割结果

序列编号	序列长度	查准率	查全率
01_01	2 751	085	0.94
01_02	4 346	0.91	0.89
01_03	4 510	0.85	0.83
01_04	4 298	0.84	0.90
01_05	4 376	0.86	0.90
01_06	5 311	0.78	0.83
01_07	4 839	0.95	0.88
01_08	4 242	0.84	0.84
01_09	3 266	0.86	0.91

如表 2 所示，除序列 01_06 外，其他序列动作的查准率和查全率均超过 0.80。平均查全率和查准率分别为 0.86, 0.88。

对文献[3]中讨论的其他 3 种算法，即 PCA,PPCA,GMM 进行比较，结果如表 3 所示。

表 3　文献[3]中 3 种算法的比较结果

方法	查准率	查全率
PCA	0.79	0.88
PPCA	0.92	0.95
GMM	0.77	0.71

文献[3]中使用的数据为人体运动捕捉数据，且 2 种衡量标准均建立在手工分割的基础上，这 3 种算法都没有使用先验知识，所以，和本文的方法可比性很强。由表 3 可知，本文方法的查准率优于 PCA 算法，且查准率和查全率比 GMM 算法高很多，但低于 PPCA 算法。

5　结束语

最大熵算法相对于其他算法具有不需要使用先验知识和易于实现的优点。其缺点主要是对数据的聚类数目约束性较强，对高维数据的信息损失量较大，它不能实时地对时间序列进行分割，如果时间序列过于平稳将导致无法分割。在下一步工作中，可以增加投票专家的数量，以提高投票准确性，或通过使用其他高维数据预处理方法有效去除不相关维度，以达到更好的分割效果。

参考文献

[1] Nurnberger G, Sommer M, Straug H. An Algorithm for Segment Approximation Numer[J]. Numer. Math., 1986, 48: 463-477.

[2] Cohen P, Heeringa P, Adams B. An Unsupervised Algorithm for Segmenting Categorical Timeseries into Episodes[C]//Proceedings of the ESF Exploratory Workshop on Pattern Detection and Discovery. Berlin, Germany: Springer, 2002: 49-62.

[3] Barbic J, Safonova A, Pan Jiayu, et al. Segmenting Motion Capture Data into Distinct Behaviors[C]//Proceedings of Graphics Interface Conference. New York, USA: ACM Press, 2004: 185-194.

[4] 李爱国, 覃　征. 在线分割时间序列数据[J]. 软件学报, 2004, 15(11): 1671-1679.

[5] 李爱国, 覃　征. 时间序列数据的稳健最优分割[J]. 西安交通大学学报, 2003, 37(4): 338-342.

编辑　陈　晖

(上接第 25 页)

参考文献

[1] Dorigo M, Maniezzo V, Colorni A. The Ant System: Optimization by a Colony of Cooperating Agents[J]. IEEE Transactions on Systems, Man, and Cybernetics, 1996, 26(1): 29-41.

[2] Dorigo M, Stützle T. Ant Colony Optimization[M]. Cambridge, USA: MIT Press, 2004.

[3] Stutzle T. Parallelization Strategies for Ant Colony Optimization[C]// Proc. of the 5th International Conf. on Parallel Problem Solving from Nature. [S. l.]: Springer-Verlag, 1998.

[4] Bullnheimer B, Kotsis G, Strauss C. Parallelization Strategies for Ant System[M]. [S. l.]: Kluwer Academic Publishers, 1998.

[5] Islam M T, Thulasiraman P, Thulasiram R K. A Parallel Ant Colony Optimization Algorithm for All-pair Routing in Manets[C]//Proc. of the 17th International Symposium on Parallel and Distributed Processing. Washington, USA: IEEE Computer Society, 2003.

[6] Randall M, Lewis A. A Parallel Implementation of Ant Colony Optimization[J]. Journal of Parallel and Distributed Computing, 2002, 62(9): 1421-1432 .

[7] Pan J S, McInnes F R, Jack M A. Application of Parallel Genetic Algorithm and Property of Multiple Global Optima to VQ Codevector Index Assignment for Noisy Channels[J]. IEEE Electronics Letters, 1996, 32(4): 296-297.

[8] Chu Shu-Chuan, Roddick J F, Pan J S, et al. Parallel Ant Colony Systems[C]//Proc. of International Symposium on Methodologies for Intelligent Systems. Aizu, Japan: [s. n.], 2003.

编辑　顾姣健