



文章编号:1001-9081(2017)12-3477-05

DOI:10.11772/j.issn.1001-9081.2017.12.3477

面向不等长多维时间序列的聚类改进算法

霍纬纲*, 程震, 程文莉

(中国民航大学 计算机科学与技术学院, 天津 300300)

(*通信作者电子邮箱 wghuo@cauc.edu.cn)

摘要:针对已有基于模型的多维时间序列(MTS)聚类算法处理不等长MTS速度较慢的问题,提出了一种基于LR分量提取的MTS聚类算法(MUTSCA(LRCE))。首先,采用等频离散化方法符号化MTS;然后,计算用于表达MTS样本各维时间序列之间时序模式的LR向量,对每个LR向量进行排序后从其两端提取固定数目的不同关键分量,所有提取的关键分量拼接形成表示MTS样本的模型向量,该过程将不等长MTS样本集转换为等长的模型向量集;最后,采用k-means算法对生成的等长模型向量集进行聚类分析。在多个公共数据集上的实验结果表明,与基于模型的MTS聚类算法——MUTSCA(LR)相比,所提算法能够在保证聚类效果的前提下,显著提高不等长MTS数据集的聚类速度。

关键词:等频离散化;k-means聚类;时序模式;多维时间序列;效率

中图分类号:TP311.13; TP181 **文献标志码:**A

Improved clustering algorithm for multivariate time series with unequal length

HUO Weigang*, CHENG Zhen, CHENG Wenli

(School of Computer Science and Technology, Civil Aviation University of China, Tianjin 300300, China)

Abstract: Aiming at the problem of slow speed of the existing model-based Multivariate Time Series (MTS) clustering algorithm when dealing with MTS with unequal length, an improved clustering algorithm named Multivariate Time Series Clustering Algorithm based on Lift Ratio (LR) Component Extraction (MUTSCA(LRCE)) was proposed. Firstly, the equal frequency discretization method was used to symbolize MTS. Then, the LR vector was calculated to express the temporal pattern between the dimensions of MTS samples. Each LR vector was sorted and a fixed number of different key components were extracted from both ends. All the extracted key components were spliced to form a model vector for representing the MTS samples. The MTS sample set with unequal length was transformed into a model vector set with equal length. Finally, the k-means algorithm was used for the clustering analysis of generated model vector set with equal length. The experimental results on multiple common data sets show that, compared with the model-based MTS clustering algorithm named Multivariate Time Series Clustering Algorithm (LR) (MUTSCA (LR)), the proposed algorithm can significantly improve the clustering speed of MTS data sets with unequal length under the premise of guaranteeing clustering effect.

Key words: equal frequency discretization; k-means clustering; temporal pattern; Multivariate Time Series (MTS); efficiency

0 引言

时间序列作为数据挖掘问题中一种重要的数据形式,是聚类领域的重点研究对象之一。从变量数目的角度看,时间序列可以分为单维时间序列和多维时间序列(Multivariate Time Series, MTS)。MTS是多维变量按照时间顺序所记录的一系列观察值的集合,广泛存在于商业、金融业、航空业、社会学、生物学等众多应用背景中,MTS数据集有其重要的数据特点:1)MTS的各维之间可能存在一些固有的时序关系,简单地拆分和降维容易造成关键信息的丢失;2)MTS样本之间长度往往不等,故与样本长度相关的预处理算法不利于后期的聚类分析;3)每个MTS样本由多个单维时间序列组成,这些时间序列的数据类型可能不同。以上特点增加了MTS数据集的聚类难度,使得面向MTS数据集的聚类算法有别于一

般的方法。

现有的时间序列聚类算法主要包含三类^[1]:基于原始数据的聚类、基于特征的聚类和基于模型的聚类。然而在这些聚类算法中,用于MTS数据集的聚类算法相对较少。文献[2]利用单维时间序列的聚类思想,给多维时间序列的各个维度赋予特定的权值,每个行向量作为一个时间点。由于MTS样本长度不等,样本之间的相似度使用动态时间弯曲(Dynamic Time Warping, DTW)度量,最佳匹配路径上每一对时间点的多维向量之间的距离利用闵可夫斯基参数模型计算。该算法需要领域知识为各个变量赋予权值,且DTW距离度量方法的计算量较大。文献[3]提出基于变量相关性的MTS特征表示方法,通过协方差反映系统中各个参数的相关关系,将MTS样本转化为协方差矩阵;MTS集所有的协方差矩阵拼接为综合协方差矩阵,对该协方差矩阵进行主成分分

收稿日期:2017-05-18;修回日期:2017-07-05。

基金项目:国家自然科学基金资助项目(61301245);中国民航联合研究基金资助项目(U1633110)。

作者简介:霍纬纲(1978—),男,山西洪洞人,副教授,博士,CCF会员,主要研究方向:数据挖掘、模糊分类;程震(1991—),男,江苏沛县人,硕士研究生,主要研究方向:数据挖掘;程文莉(1992—),女,河南鹤壁人,硕士研究生,主要研究方向:大数据。



析得到各 MTS 的特征矩阵。该方法可以将数值型不等长 MTS 数据集转变为大小相同的特征矩阵集合, 处理结果可用于聚类分析。文献[4]提出了一种基于参数交互关系的 MTS 聚类方法, 指出 MTS 中的任一维变量都可以被其他解释变量近似线性组合表示, 且将一维线性关系纳入了考虑范畴, 假定这些变量间的线性相关关系可以用来进行聚类, 其不足之处在于模型计算时间会随着样本数量变大而增加, 也不能处理非数值型变量。文献[5]将每一维时间序列转化为一个统计特征数组, MTS 样本由各维变量统计特征数组拼接成的向量来表示。该算法可以处理不等长时间序列, 但要求各维选取的统计特征必须一致导致其在处理混合型 MTS 数据集时会遇到困难。文献[6]针对 MTS 数据集存在的样本之间不等长、数据类型多样和噪声等问题, 提出了一种基于协方差矩阵与测地线距离 (geodesic-based distance) 的 MTS 聚类算法。该算法首先将 MTS 样本转化为协方差矩阵; 然后将协方差矩阵从黎曼空间映射到欧氏空间; 最后对矩阵集进行聚类。如果使用基于距离的聚类算法, 上述映射过程可以省略, 协方差矩阵之间的距离度量方法使用测地线距离。Zhou 等^[7]提出了一种基于模型的多维时间序列聚类算法——MUTSCA〈LR〉(Multivariate Time Series Clustering Algorithm 〈Lift Ratio〉), 该聚类算法假设目标数据集由一系列概率分布模型系统生成, 不同的系统将生成相异的多维时间序列。该算法先将连续型数值符号化; 然后在符号化样本上计算由 LR(Lift Ratio) 向量表示的时序模式, 将时序模式累加生成用来表示 MTS 样本的模型向量; 最后对模型向量集进行聚类。它不需要特定的领域知识, 同时可以处理包含数值和非数值型变量的混合型 MTS 数据集。但实际应用中的 MTS 数据集, 各样本的长度往往不等, 该算法在处理不等长多维时间序列时需要利用 DTW^[8] 或滑动窗口^[9] 来度量模型向量之间的相似性, 造成算法的时间开销较大。

针对上述算法在处理 MTS 数据集时的不足, 本文将在 MUTSCA〈LR〉的基础上进行改进, 提出一种基于 LR 分量提取的多维时间序列聚类算法 (Multivariate Time Series Clustering Algorithm based on LR Component Extraction, MUTSCA〈LRCE〉)。该算法使用等频离散化方法对 MTS 数据集进行符号化; 在符号化后的 MTS 样本每一维时间序列之间计算时序模式向量 LR, 对 LR 向量进行排序并提取向量两端固定数目不同数值的分量, 同一 MTS 样本内的提取分量拼接成一个模型向量 (Model Vector, MV), MTS 数据集由 MV 向量的集合表示; 最后对模型向量集使用 k-means 算法进行聚类分析。实验结果表明, 改进算法可以快速地完成不等长 MTS 数据集的聚类分析。

1 相关知识背景

1.1 符号表示与相关定义

定义 1 多维时间序列集。一个包含 N 个多维时间序列样本的多维时间序列集可表示为 $S = \{S^1, S^2, \dots, S^N\}$ 。若每个样本 S^i 包含 m 个变量, 每一维时间序列长度为 n , 则 S^i 可表示为 $S^i = \{S_1^i, S_2^i, \dots, S_m^i\}$, 其中第 k 维时间序列为 $S_k^i = \{S_{k,1}^i, S_{k,2}^i, \dots, S_{k,n}^i\}$ 。文中符号化后的多维时间序列集则表示为 $S^* = \{S^{1*}, S^{2*}, \dots, S^{N*}\}$ 。

定义 2 时序模式。用 LR 向量表示的时序模式的计算方法^[6]如下:

$$\text{Lift}(S_{j,t-\tau}^{i*} \Rightarrow S_{j,t}^{i*}) = \frac{\Pr(S_{j,t}^{i*} | S_{j,t-\tau}^{i*})}{P(S_{j,t}^{i*})} = \frac{\text{freq}(S_{j,t-\tau}^{i*} \cap S_{j,t}^{i*})}{\text{freq}(S_{j,t}^{i*}) * \text{freq}(S_{j,t-\tau}^{i*})}; 1 \leq \tau < t \quad (1)$$

其中: τ 是常数, 表示时间偏移量; $S_{j,t}^{i*}$ 与 $S_{j,t-\tau}^{i*}$ 分别表示第 i 个样本的第 j 维时间序列在 t 和 $t - \tau$ 时刻所对应的符号值; $P(S_{j,t}^{i*})$ 表示 $S_{j,t}^{i*}$ 的发生概率; $\Pr(S_{j,t}^{i*} | S_{j,t-\tau}^{i*})$ 表示当 $S_{j,t-\tau}^{i*}$ 发生时 $S_{j,t}^{i*}$ 发生的条件概率; $\text{freq}(S_{j,t}^{i*})$ 表示 $S_{j,t}^{i*}$ 在时间序列 S_j^{i*} 中出现的次数; t 是时间变量, 与 MTS 长度 n 有关, 因此该公式将生成一个长度为 $n - \tau$ 的向量, 称为 Lift Ratio(LR)。它反映了每个时刻 $t - \tau$ 的 $S_{j,t-\tau}^{i*}$ 对 $S_{j,t}^{i*}$ 的依赖程度, 以此反映样本 S^{i*} 的变量 j 的时间序列 S_j^{i*} 内部的时序模式。

式(1)反映的是同一维时间序列 S_j^{i*} 内部的 LR 计算公式, 当式(1)中的 $S_{j,t}^{i*}$ 换成 $S_{k,t}^{i*}$ ($k \neq j$) 时, 则表示不同维的时间序列 S_j^{i*} 与 S_k^{i*} 之间的 LR 计算公式。同一变量序列内部计算的 LR 称为 *intra-pattern*, 不同变量时间序列之间的 LR 称为 *inter-pattern*。

1.2 MUTSCA〈LR〉算法

MUTSCA〈LR〉算法首先使用 MDD (Mode-Driven Discretization) 算法^[10]对 MTS 样本进行符号化, MDD 符号化过程如下: MDD 算法将 MTS 作为由若干个向量组成的矩阵, 行为向量, 列为变量维; 首先选取 MTS 中独立冗余度 (Multiple interdependence Redundancy, MR) 值最大的变量维作为标签属性, 标签维每一行的符号值为该行向量的分类标签; 然后利用有监督的离散化算法 OCDD (Optimal Class-Dependent Discretization)^[11] 对其他数值型变量维进行符号化。在符号化后的 MTS 样本上利用式(1)计算时序模式 LR 并累加得到模型向量。最后, 利用 k-means 算法对模型向量集进行聚类分析。MUTSCA〈LR〉算法模型向量计算方法如下:

```

输入  $S^* = \{S^{1*}, S^{2*}, \dots, S^{N*}\}$ , 其中  $S^{i*} = \{S_1^{i*}, S_2^{i*}, \dots, S_m^{i*}\}$ ;
输出  $N$  个模型向量组成的向量集  $\text{finalLR}$ 。
For 每个多维时间序列  $S^{i*}$ 
    For 每个变量  $S_j^{i*}$ 
        利用式(1)计算 intra-pattern;
        For 每个变量  $S_k^{i*}$  ( $k \neq j$ )
            利用式(1)计算 inter-pattern;
             $\text{finalLR}_i += \text{inter-pattern}$ ;
        End
         $\text{finalLR}_i += \text{intra-pattern}$ ;
    End
     $\text{finalLR} = \text{finalLR} \cup \{\text{finalLR}_i\}$ ;
End

```

MUTSCA〈LR〉算法存在两个问题:

1) MDD 算法中采用的 OCDD 利用动态规划的思想选取分割点集, 若处理长时间序列需要较多的离散化符号, 则要求候选分割点集包含较多的元素, 导致 OCDD 的开销过高, 离散化执行效率低下。

2) 由于 $j, k \in \{1, 2, \dots, m\}$, 每个样本计算得到 m^2 个时序模式向量 LR, 多维时间序列 S^{i*} 最终生成的时序模型 finalLR_i 由 m^2 个 LR 向量累加求和得到。如果多维时间序列集 S^* 中的各个样本 S^{i*} 长度不等, 则 MUTSCA〈LR〉生成的模型向量集 finalLR 中的各个 finalLR_i ($i \in \{1, 2, \dots, N\}$) 长度也不相等。上述问题增加了相似性度量的难度, 造成该算法



在聚类过程的耗时较长。

2 MUTSCA〈LRCE〉算法

2.1 多维时间序列的等频离散化

针对 MDD 算法效率较低的问题,文中采用等频离散化(Equal Frequency Discretization, EFD) 算法进行符号化。EFD 是一种简单的离散化方法,它需要事先给定一个参数来决定离散化后最终的离散符号个数,记为 num_bin 。在没有领域知识的情况下,各个样本的 num_bin 往往难以确定。传统的等频离散化方法需要用户随机给出 num_bin 的取值,如此会导致 MTS 聚类分析的效果不稳定。 num_bin 的取值同 MTS 样本长度有关,但 MTS 数据集各样本长度差异过大将导致 num_bin 的值域范围较宽。鉴于以上原因,本文基于样本长度和变异系数提出一个用来为样本 $S^i (i \in \{1, 2, \dots, N\})$ 选取 num_bin 值的计算方法:

$$num_bin = (\text{int}) C * \sqrt{n}; num_bin \leq n/2 \quad (2)$$

其中: C 为由 MTS 样本集 S 确定的唯一常数, n 是 S^i 的时间序列长度。输入样本集 $S = \{S^1, S^2, \dots, S^N\}$, 统计 S 中各 MTS 长度的均值 M 和标准差 V_e , 计算变异系数 $C_v = V_e/M$, C 为变异系数的倒数 $1/C_v$ 。从式(2) 可知, 当 MTS 样本集各样本长度的 C_v 取值越大, 参数 C 的取值越小, 从而缓解各 MTS 样本离散符号数目差异大的问题。文中的等频离散化方法具体描述如下。

```

输入 样本集  $S = \{S^1, S^2, \dots, S^N\}$ , 其中  $S^i = \{S_1^i, S_2^i, \dots,$ 
 $S_m^i\}$ ;
输出 符号化后的样本集  $S^* = \{S^{1*}, S^{2*}, \dots, S^{N*}\}$ 。
统计样本集各个时间序列的均值和方差,并计算系数  $C$ ;
For 样本集每一个多维时间序列  $S^i$ 
    利用式(2)计算  $S^i$  的  $num\_bin$ ;
    For 所有变量序列  $S_j^i (j \in \{1, 2, \dots, m\})$ ;
        根据  $num\_bin$  的取值,对该  $S_j^i$  进行等频离散化,得到  $S_j^{i*}$ ;
         $S^{i*} = S^* \cup S_j^{i*}$ ;
    END
     $S^* = S^* \cup S^{i*}$ ;
END

```

2.2 模型向量集的生成方法及其聚类

首先在样本 S^{i*} 每一对变量的时间序列 $(S_{v1}^{i*}, S_{v2}^{i*}) (v1, v2 \in \{1, 2, \dots, m\})$ 上计算时序模式 LR 向量;然后将 LR 向量进行排序并选取排序后向量两端 K 个不同数值的分量, 数值较大的分量反映了时序模式的主要特征, 数值较小的分量反映了时序模式的特殊状态;最后 S^* 的全部提取分量将组成模型向量 MV_i 。根据 $v1, v2$ 的取值, 样本 S^{i*} 利用式(1) 计算出 m^2 个 LR 向量, 每个 LR 向量中提取 K 个分量, 因此 MUTSCA〈LRCE〉生成的模型向量 MV_i 长度 L 仅与分量个数 K 、参数个数 m 相关, $L = m^2 * K$ 。模型向量集生成方法描述如下。

```

输入 离散化后的样本集  $S^* = \{S^{1*}, S^{2*}, \dots, S^{N*}\}$ ,
 $S^{i*} = \{S_1^{i*}, S_2^{i*}, \dots, S_m^{i*}\}$ , 分量个数  $K$ ;
输出 模型向量集  $MV = \{MV_1, MV_2, \dots, MV_N\}$ 。
For 每一个多维时间序列  $S^i$ 
    For 多维时间序列的所有参数  $v1 (1 \leq v1 \leq m)$ 
        For 多维时间序列的所有参数  $v2 (1 \leq v2 \leq m)$ 
            利用式(1)计算 LR 向量;
            对 LR 进行排序,从中提取首尾  $K$  个数值不同的分量,并

```

拼接到 MV_i ;

End

End

$MV = MV \cup MV_i$;

End

文中采用 k -means 算法对模型向量集 MV 进行聚类分析,首先随机选取 k 个模型向量作为初始簇中心,计算模型向量与各簇中心的欧氏距离,并将该向量分配给最相似的簇, MV 分类完毕后更新簇中心。重复上述步骤,直至分簇结果不再变化,输出分簇结果向量 $KM = [km_1, km_2, \dots, km_N]$, 其中 $km_i (km_i \in \{1, 2, \dots, k\})$ 表示样本 S^i 的分簇编号。聚类过程如下。

```

输入 模型向量集  $MV = \{MV_1, MV_2, \dots, MV_N\}$ , 分簇个数  $k$ ;
输出 分簇结果向量  $KM = [km_1, km_2, \dots, km_N]$ 。
从  $MV$  中随机选取  $k$  个向量作为初始簇中心集合  $Core = \{core_1,$ 
 $core_2, \dots, core_k\}$ ;
构建一个临时簇中心集合  $Core' = \{core'_1, core'_2, \dots, core'_k\}$ ;
构建一个数组  $Num[k]$ ;
初始化  $KM = [0, 0, \dots, 0]$ ;
Do
    初始化  $Core'$  为 0 向量集,  $Num[k]$  为 0 向量;
    For 模型向量  $MV_i (1 \leq i \leq N)$ 
        计算每个模型向量与  $Core$  的  $k$  个簇中心的相似度,并将它分
        配到最相似的簇  $\theta$ ,其中  $\theta \in \{1, 2, \dots, k\}$ ;
         $km_i = \theta$ ;
         $core'^{km_i} = MV_i$ ;
         $Num[km_i - 1]++$ ;
    End
    利用  $Core'$  与  $Num[k]$  更新簇中心,结果返还给  $Core$ ;
Until  $KM$  no change

```

3 实验与结果分析

算法使用 Java 进行编程实现,实验在一台配备 Intel 四核 3.80 GHz 处理器、4 GB 内存、装有 Window 7 系统的 PC 上进行。

3.1 实验准备与数据集介绍

选用 4 个来自 UCI 的 MTS 数据集: EMCPAD (EMG Physical Action Data set)、EMGLL (EMG dataset in Lower Limb)、AReM (Activity Recognition system based on Multisensor data fusion)、DSAD (Daily and Sports Activities Data set), 其中 AReM 与 DSAD 为等长 MTS 数据集,详见表 1。

1) EMCPAD。样本数目 80, 包括 3 位男性和 1 位女性实验者。每个实验者做 20 个动作(20 个样本),包括 10 个攻击性动作和 10 个一般动作,样本长度大多在 10 000 左右。实验使用动作的性质为标签,即攻击性和非攻击性。

2) EMGLL。样本数目 66, 包括 11 位膝关节患者和 11 位正常人。每个实验者做 3 种运动,各样本长度波动较大。实验使用自然人的分类作为标签,即患者和正常人。

3) AReM。样本数目 87, 包括 7 个类型一的弯腰动作、5 个类型二的弯腰动作、15 个骑车动作、15 个躺动作、15 个坐动作、15 个站立动作、15 个走路动作,各样本长度均为 480。实验使用动作的类型作为标签,共 7 种类别。该数据集是为了验证改进算法 MUTSCA〈LRCE〉在等长 MTS 数据集上可以维持 MUTSCA〈LR〉的聚类效果。



4) DSAD。样本数目 9 120,由 8 位实验参与者完成 19 种动作,每个动作包含 60 个样本。该数据集旨在检验算法 MUTSCA〈LRCE〉在大样本集下的聚类效果。

实验过程中参数设置如下,LR 计算的时间延时 τ 的值取 5, k -means 聚类算法的簇中心个数 k 取值为数据集的类别个数,提取排序后 LR 向量的首尾分量个数 K 取 10。由于实验数据集有标签,故采用 F-measure 和信息熵作为实验中多维时间序列聚类算法的评价指标。

表 1 实验数据介绍
Tab. 1 Introduction to experimental data

数据集	变量数	样本数	序列最大长度	序列最小长度	类别数
EMGPAD	8	80	15 000	8 000	2
EMGLL	5	66	40 000	5 000	2
AReM	6	87	480	480	7
DSAD	45	9 120	125	125	19

为了便于算法 MUTSCA〈LR〉与 MUTSCA〈LRCE〉对比,本实验使用两种方法对 MUTSCA〈LR〉进行修改使其能够处理不等长 MTS 数据集:1) 使用 DTW 计算不等长模型向量之间的距离,采用文献[12]提出的基于 DTW 的全局平均法进行 k -means 聚类中心点的更新;2) 使用滑动窗口计算不等长模型向量之间的距离,簇中心点的更新方法如下:假设某次 k -means 迭代过程产生的簇中心集合为 $Core = \{core_1, core_2, \dots, core_k\}$, $length(core_i)$ 表示向量 $core_i$ 的长度。设 $core_i$ 所在簇中有 n_i 个模型向量,在簇中心 $core_i$ ($1 \leq i \leq k$) 的更新过程中,创建两个长度为 $length(core_i)$ 的 0 向量,记为 Sum 和 $weight$ 。对于 $core_i$ 簇中的某模型向量 MV_j ($1 \leq j \leq n_i$),若 $length(MV_j) \geq length(core_i)$, 使用滑动窗口在 MV_j 上截取与 $core_i$ 最相似的子序列累加至 Sum 、 $weight$ 各分量数值加 1;若 $length(MV_j) < length(core_i)$, 使用滑动窗口在 $core_i$ 上找到与 MV_j 最相似子序列的起始位置 st , 将 MV_j 累加到 $Sum[st]$ 至 $Sum[(st + length(MV_j) - 1)]$,且 $weight$ 中相应位置的分量加 1。更新后簇中心 $core'_i$ 的计算公式为: $core'_i[t] = Sum[t]/weight[t]$ ($0 \leq t < length(core_i)$)。

3.2 本文的等频离散化方法评估

采用包含较长 MTS 样本的数据集 EMGPAD 和 EMGLL 对本文 EFD 与 MDD 算法进行评估。两种离散化方法的处理结果均使用上述基于滑动窗口的 MUTSCA〈LR〉算法进行聚类分析。根据文献[10]的经验值,算法 MDD 对每个样本 S^i 的分割点候选集元素个数取值为 $\sqrt{length(S^i)/3}$,其中 $length(S^i)$ 表示样本 S^i 的长度。本文测试了该参数与 MDD 符号化性能之间的关系,设实验选用的候选分割点集包含元素个数 $y = \theta * \sqrt{length(S^i)/3}$,计算结果进行“上取整”处理作为实际选取元素个数。这里用系数 θ 表示候选分割点集的大小,实验结果如图 1 所示,图 1 中横坐标表示候选分割点集大小 θ 。由图 1 可以看出,随着候选分割点集元素个数增加,算法 MDD 得到的符号化 MTS 样本集聚类评价提高,但离散化执行时间会迅速增长。

表 2 给出了算法 EFD 与 MDD 的测评结果,对于样本 S^i ,MDD 使用的候选分割点集元素个数取值为 $\sqrt{length(S^i)/3}$ 。从聚类评价指标值可以看出,样本集 EMGPAD 与 EMGLL 使用 EFD 离散化方法所得到的符号化样本集在进行聚类分析时效果较好,其 F-measure 值分别为 0.896 与 0.904,而 MDD

算法相应的结果为 0.870 与 0.742。其中,MDD 在数据集 EMGLL 上符号化效果较差的现象与候选分割点集包含元素数目不足有关,但由图 1 的实验结果可知,通过增加候选分割点集元素个数来提高聚类评价结果的方法时间代价过高。从离散化时间的角度看,本文 EFD 算法的执行时间分别为 4 065 ms 和 5 348 ms,明显低于 MDD 算法的执行时间。

表 2 等频离散化与 MDD 效果评估
Tab. 2 Evaluation of equal frequency discretization and MDD

算法	数据集	运行时间 (离散化)/ms	F-measure	熵
EFD	EMGPAD	4 065	0.896	0.483
	EMGLL	5 348	0.904	0.405
MDD	EMGPAD	254 899	0.870	0.507
	EMGLL	590 204	0.742	0.842

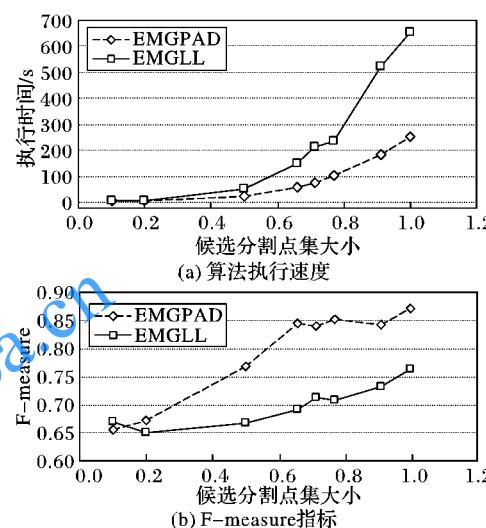


图 1 MDD 算法评估

Fig. 1 Evaluation of MDD algorithm

3.3 MUTSCA〈LRCE〉算法评估

使用 4 个 MTS 数据集对改进算法 MUTSCA〈LRCE〉进行评估,如表 3 所示,其中 MUTSCA〈LR〉使用 MDD 进行样本集符号化,MUTSCA〈LRCE〉则采用改进的 EFD 算法。聚类时间包括模型计算时间和 k -means 执行时间,由表 3 可知,在不等长 MTS 数据集 EMGPAD 与 EMGLL 上,改进后的 MUTSCA〈LRCE〉因包含向量排序过程,故时间略长,这里认为算法改进前后的模型计算时间基本一致。算法 MUTSCA〈LRCE〉的 k -means 执行时间为 35 ms 和 16 ms,而原算法 k -means 的执行时间明显较长。这是由于算法 MUTSCA〈LRCE〉生成的模型向量长度仅取决于 MTS 参数个数 m 以及分量个数 K ,模型向量长度固定,使得算法聚类速度较快。此外,基于滑动窗口的 MUTSCA〈LR〉在数据集 EMGPAD 上执行时间少于 EMGLL 的原因是 EMGPAD 中各样本长度波动较小,模型向量之间的相似性度量包含的窗口滑动次数较少。从聚类结果来看,MUTSCA〈LRCE〉与 MUTSCA〈LR〉在 EMGPAD 上的聚类效果相当,而在数据集 EMGLL 上算法 MUTSCA〈LR〉的 F-measure 值在 0.75 附近,由实验 3.2 节的分析可知其在该数据集上失效的原因在于该数据集存在长度较短的样本,它们的候选分割点集中元素较少,导致 MDD 离散化效果不佳,影响了后续的聚类分析。

数据集 AReM 是样本数目较少的等长多维时间序列数据集,由于样本长度较短,聚类时间相互之间区分度低,参考价



值不大。从聚类结果的角度看,三种算法的熵和 F-measure 值基本一致。数据集 DSAD 为包含较多等长多维时间序列样本,该部分实验选取了 4800 个样本,由于本文算法生成的模型向量维度与变量个数有关,模型向量维度 L 大于样本长度,其中 $L = m^2 * K$, 样本长度为 125。因此模型向量维度高,聚类时间较长。同时,较高维度的模型向量包含更多特征,故聚类效果优于算法 MUTSCA (LR)。所以,算法 MUTSCA (LRCE) 在多标签的等长 MTS 数据集上仍然有效。

表 3 MUTSCA (LRCE) 的聚类工作评估

Tab. 3 Clustering work evaluation of MUTSCA (LRCE)

算法	数据集	模型计算时间/ms	k-means 执行时间/ms	F-measure	熵
基于滑动窗口的	EMCPAD	33 922	10 044	0.869	0.506
	EMGLL	22 887	134 217	0.758	0.788
MUTSCA (LR)	AReM	462	6	0.655	1.450
	DSAD	335 406	254	0.452	1.911
基于 DTW 的	EMCPAD	32 732	113 741	0.860	0.532
	EMGLL	23 454	275 008	0.750	0.809
MUTSCA (LR)	AReM	457	447	0.667	1.432
	DSAD	321 756	23 007	0.457	1.713
MUTSCA (LRCE)	EMCPAD	36 718	35	0.888	0.451
	EMGLL	23 196	16	0.889	0.495
	AReM	891	13	0.655	1.452
	DSAD	377 473	15 351	0.822	0.225

为了评估聚类簇个数 k 值对 k-means 聚类结果的影响,本文从数据集 DSAD 中选取两组样本子集进行实验,实验结果如表 4 所示。第一组样本子集包含样本数目 240 个,分别来自 4 种不同类型动作的样本集,标签数目为 4。第二组样本子集包含 1920 个样本,实验以同一动作类型的同一个行为主体的 60 个样本为一组,抽取 32 组,所以样本子集标签数目为 32。

表 4 MUTSCA (LRCE) 算法 k 值对聚类结果的影响Tab. 4 Influence of k value in MUTSCA (LRCE) algorithm on clustering results

第一组样本子集		第二组样本子集	
k	熵	k	熵
1	2.000	10	2.314
2	1.001	15	2.021
3	0.992	20	1.213
4	0.495	25	1.210
5	0.470	32	1.014
6	0.285	40	0.674
7	0.308	45	0.671
8	0.271	50	0.729

由表 4 可以看出,当 k 值小于标签个数时聚类效果较差,当参数 k 略大于标签数时,聚类效果较好。这是因为 k 值较小时,不同标签的样本容易被合并到同一类簇中。当 k 值略大于样本标签数目时, k-means 分簇更精细,噪声样本对聚类过程的影响降低。

4 结语

本文基于时序模式 Lift Ratio 向量的 MTS 表示方法提出了 MUTSCA (LRCE) 算法,该算法利用改进的等频离散化方法对 MTS 进行符号化,通过 LR 向量分量提取的方式将不等长

MTS 样本转化为等长的模型向量。实验结果表明本文算法可以更好地对不等长 MTS 数据集进行聚类分析。时间序列是数据之间具有严格上下文关系的一类特殊数据对象,LR 向量所展现的 MTS 时序模式仅反映了时间点之间的时序关系,如何利用时间段之间的时序关系进行 MTS 聚类并减少时间段处理过程中造成的信息丢失有待进一步研究。

参考文献 (References)

- [1] LIAO T W. Clustering of time series data — a survey [J]. Pattern Recognition, 2005, 38(11): 1857 – 1874.
- [2] CHANDRA B, GUPTA M, GUPTA M P. A multivariate time series clustering approach for crime trends prediction [C]// Proceedings of the 2008 IEEE International Conference on Systems, Man & Cybernetics. Piscataway, NJ: IEEE, 2008: 892 – 896.
- [3] 李海林. 基于变量相关性的多元时间序列特征表示[J]. 控制与决策, 2015, 30(3): 441 – 447. (LI H L. Feature representation of multivariate time series based on correlation among variables [J]. Control and Decision, 2015, , 30(3): 441 – 447.)
- [4] PLANT C, WOHL SCHLAGER A M, ZHERDIN A. Interaction-based clustering of multivariate time series [C]// Proceedings of the 9th IEEE International Conference on Data Mining. Washington, DC: IEEE Computer Society, 2009: 914 – 919.
- [5] WANG X Z, WIRTH A, WANG L. Structure-based statistical features and multivariate time series clustering [C]// Proceedings of the 2007 IEEE International Conference on Data Mining. Piscataway, NJ: IEEE, 2007: 351 – 360.
- [6] SUN J. Clustering multivariate time series based on Riemannian manifold [J]. Electronics Letters, 2016, 52(19): 1607 – 1609.
- [7] ZHOU P Y, CHAN K C C. A model-based multivariate time series clustering algorithm [C]// Proceedings of the 2014 International Workshops Trends and Applications in Knowledge Discovery and Data Mining, LNCS 8643. Berlin: Springer, 2014: 805 – 817.
- [8] KEOGH E. Exact indexing of dynamic time warping [J]. Knowledge and Information Systems, 2005, 7(3): 358 – 386.
- [9] YE L, KEOGH E. Time series shapelets: a new primitive for data mining [C]// KDD 2009: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York: ACM, 2009: 947 – 956.
- [10] WONG A K C, WU B, WU G P K, et al. Pattern discovery for large mixed-mode database [C]// CIKM 2010: Proceedings of the 19th ACM International Conference on Information & Knowledge Management. New York: ACM, 2010: 859 – 868.
- [11] LIU L, WONG A K C, WANG Y. A global optimal algorithm for class-dependent discretization of continuous data [J]. Intelligent Data Analysis, 2004, 8(2): 151 – 170.
- [12] PETITJEAN F, KETTERLIN A, GANCARSKI P. A global averaging method for dynamic time warping, with applications to clustering [J]. Pattern Recognition, 2011, 44(3): 678 – 693.

This work is partially supported by the National Natural Science Foundation of China (61301245), the Joint Funds of Civil Aviation Administration of China (U1633110).

HUO Weigang, born in 1978, Ph. D., associate professor. His research interests include data mining, fuzzy clustering.

CHENG Zhen, born in 1991, M. S. candidate. His research interests include data mining.

CHENG Wenli, born in 1992, M. S. candidate. Her research interests include big data.