面向临床检验指标的非同步时间序列聚类算法研究

陈德华[1] 韩学士[1] 乐嘉锦[1] 朱立峰[2]
(东华大学计算机科学与技术学院 上海 200051)[1]
(上海交通大学医学院附属瑞金医院计算机中心 上海 200025)[2]

摘 要 对临床检验指标时间序列进行聚类,从中发现临床检验指标变化趋势相似的患者群体,对开展精准医疗具有非常重要的价值。考虑到不同患者的检验次数及检验时间点不完全同步,首先通过对非同步时间序列进行预处理,实现不同时间序列维度及时间点的同步化。在此基础上,通过引入一个用户自定义参数即噪声点占有率 NoisePro,对 DBScan 算法进行改进,提出了一种基于密度划分思想的非同步临床检验指标时间序列聚类 LabTS-CLU 算法。最后利用某三甲医院十余万糖尿病患者近 10 年的糖化血红蛋白时间序列数据集进行实验,结果证明了所提算法的有效性。

关键词 临床检验指标,非同步时间序列,密度聚类
中图法分类号 TP301.5 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.9.023

Efficient and Effective Clustering Algorithm for Asynchronous Time Series of Clinical Laboratory Indicators

CHEN De-hua[1] HAN Xue-shi[1] LE Jia-jin[1] ZHU Li-feng[2]
(School of Computer Science and Technology, Donghua University, Shanghai 200051, China)[1]
(Computer Center of Ruijin Hospital, Shanghai Jiaotong University School of Medicine, Shanghai 200025, China)[2]

Abstract Clustering for asynchronous time series of clinical laboratory indicators, and finding the patient group with similar variation trends of clinical laboratory indicators, have a very important value for the conduct of precision medicine. Taking into account the frequency of inspection and the testing time points of different patients are not fully synchronized, asynchronous time series were preprocessed to achieve the synchronization of different time dimensions and time points. On this basis, we improved the DBScan algorithm by introducing a user-defined parameter namely noises share NoisePro. Then, we proposed a LabTS-CLU time series clustering algorithm of asynchronous clinical test indicators based on density divided thoughts. Finally, experimental results on the time series of glycated hemoglobin dataset of more than 100 thousand diabetics in the past 10 years from a hospital demonstrate the effectiveness of the proposed algorithm.

Keywords Clinical indicators, Asynchronous time series, Density clustering

1 引言

临床检验指标是临床医生对患者疾病状况进行判断的重要依据,例如血糖和糖化血红蛋白就是反映糖尿病患者血糖水平的两项重要临床检验指标[1]。通过将患者屡次的临床检验指标与治疗时间关联,可建立患者临床检验指标的时间序列,这种时间序列从某种程度上反映了患者疾病状况随治疗时间的变化趋势。因此,对临床检验指标时间序列进行分析,可以揭示患者疾病变化和发展的内在规律,对临床决策和辅助诊断具有重要的现实意义。

对于慢性病如糖尿病、甲亢、肿瘤等,临床检验指标的跟踪监测及分析是一项重要的研究课题,受到医学界的广泛关注[2]。以糖尿病为例,在治疗过程中,临床医生经常希望能对糖尿病患者的临床检验指标时间序列(如血糖和糖化血红蛋白)进行聚类分析,从中发现某些指标具有相同变化趋势的患者群体。这类患者群体的发现,有助于总结出同类患者的共同性,更有助于临床医生从糖尿病患者群体的疾病发展趋势来预测患者自身的疾病进展状况,从而为制定正确的治疗方案提供参考依据。

本文以慢性病特别是糖尿病患者群体分类为目标,通过对临床检验指标的非同步时间序列建模,提出了一种针对非同步临床检验指标时间序列的聚类分析 LabTS-CLU 算法。该算法首先对非同步临床检验指标时间序列实现了同步化,包括不同时间序列的维度同步化和时间点同步化两种处理。然后,在同步化的临床检验指标时间序列的基础上,通过引入噪声点占有率 NoisePro 参数对 DBScan 算法[3]进行改进,提

到稿日期:2015-07-04 返修日期:2015-08-13
陈德华(1976—),副教授,主要研究方向为数据库、数据仓库与智慧医疗;韩学士(1990—),硕士,主要研究方向为大数据、数据库,E-mail:312769143@qq.com;乐嘉锦(1951—),教授,博士生导师,主要研究方向为数据库与数据仓库、大数据;朱立峰(1976—),高级工程师,主要研究方向为医院信息化与数据库。

出了针对非同步临床检验指标时间序列的有效聚类算法。

本文第 2 节介绍了相关工作；第 3 节给出了相关描述和定义；第 4 节阐述了非同步临床检验指标时间序列的同步化处理过程；第 5 节给出了非同步时间序列聚类算法的具体实现；最后通过一组真实数据进行了对比实验，结果证明了该聚类算法的有效性。

2 相关工作

时间序列的聚类分析一直是数据挖掘研究中的一个重要课题。时间序列聚类分析大致可分为两类：子序列聚类分析和整个时间序列聚类分析[4]。其中，子序列聚类主要基于个体时间序列滑动窗口的提取，主要目的在于发现个体时间序列不同时间窗口子序列之间的相似以及不同之处。整个时间序列聚类分析则是指对个体时间序列不进行分割，而是直接根据个体之间的相似性进行类别的集聚。本文的临床检验指标时间序列聚类是基于每位患者对应的整个时间序列进行聚类，因此这里仅对整个时间序列的聚类分析现状进行介绍。

在整个时间序列聚类分析中，以往的研究大多关注于同步时间序列的聚类算法研究。文献[5]提出分割时间序列的方法，即将完整的时间序列进行规定的切割之后，对其子序列进行建模聚类；文献[6]提出了使用 OLS 算法实现对在线时间序列的分割，大量实验证明 OLS 算法能够有效地在线监测出数据挖掘应用中感兴趣的关键变化点，从而为分析两条时间序列是否具有相似性提供关键性的支持；文献[7]提出了将时间序列进行线性划分的方法，即利用线段来近似表示时间序列，从而获取时间序列的变化趋势，这一研究工作让后来的研究工作者产生了如何通过降维的时间序列最大程度地保留原时间序列的信息的想法；文献[8]对文献[7]提出的线性划分方法进行了详细的评述以及扩充。

近些年，人们越来越重视医学领域内时间序列数据的运用与学习[9]，并且提出了众多的时间序列处理手段和聚类方法。诸多的方法基本可以以数据的类型进行不同的处理，其中最典型的时间序列数据类型是医学领域内的信号数据。文献[10]使用支持向量机和隐马尔科夫模型成功处理了多通道近红外光谱仪信号数据；文献[11]使用基于支持向量机的聚类和分类算法识别了脑负荷运转领域内的时间变化序列。

3 相关描述

本节给出将要用到的一些概念的形式化描述。

定义 1(临床检验指标) 在医学界，针对患者的某种疾病均有衡量其是否处于正常状态的一种或几种标准，这些衡量标准称为临床检验指标。例如，糖化血红蛋白(A1C)和空腹血糖(GLU)是衡量糖尿病患者疾病程度的重要指标。

临床检验指标按照检验结果值可分为数值型和非数值型两类。例如一位患者的一次血检结果中，糖化血红蛋白(HbA1C)=7.1%，白细胞=阴性，其中的糖化血红蛋白为数值型指标，而白细胞为非数值型指标。本文仅涉及数值型指标时间序列的分析与处理。

定义 2(临床检验指标时间序列) 将患者多次临床检验的某项指标值 v_i 与其检验时间(t_i)关联，可建立该患者的这项临床检验指标时间序列 $TS=(\langle v_1,t_1\rangle,\cdots,\langle v_i,t_i\rangle,\cdots,\langle v_n,t_n\rangle)$。其中，为更直观地观察分析所有患者指标间的共性，将患者的第一次检验时刻点取为起始时间，记为第 0 天，第二个检验时刻点为第二次检验时间与起始时间的间隔天数，随后依次类推，可算出每次检验的时刻点，记为 $T=\{t_1,t_2,\cdots,t_n\}$。例如，表 1 所列为一个糖尿病患者 A 的 HbA1C 指标时间序列。

表 1 患者 A 的 HbA1C 指标时间序列

v	6.5	7.1	7.9	6.8	7.3	10.7	9.4
t	0	98	210	479	694	924	1099

为了更形象地刻画患者之间的相似性与差异性，将患者指标值时间序列表示成曲线，这不仅可以直观地展现患者某个指标检查序列之间是否具有相似性，更有利于考察聚类结果的优劣。图 1 示出了某位患者 HbA1C 指标值序列曲线。

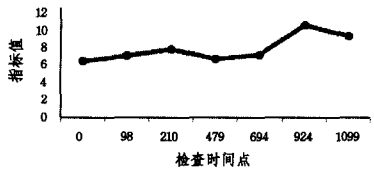

图 1 患者 A 的 HbA1C 指标序列曲线

不同患者的临床指标检验时间不完全同步，这表现在两条时间序列的各检验时间点不完全一致。例如，表 2 所列为患者 B 的 HbA1C 指标时间序列，图 2 所示为该患者的 HbA1C 时间序列曲线。

表 2 患者 B 的 HbA1C 指标值时间序列

v	8.5	6.0	5.9	6.8	6.4	6.9	7.2	6.7
t	0	18	39	67	102	145	247	439

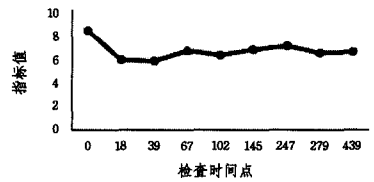

图 2 患者 B 的 HbA1C 指标序列曲线

由上述两位患者的 HbA1C 指标时间序列曲线可以看出，A、B 两个病人的时间序列的时间点是不一致的，换言之，两条时间序列是非同步的。

4 不同时间序列的同步化

针对无法直接计算非同步的临床检验指标时间序列两两之间相似性的问题，首先对时间序列进行同步化处理。临床检验指标的不同步表现在两处：1)时间序列的维度(即时间序列的对象点数量)差异性较大，因为每位患者到医院检查的天数、周期不同，导致患者之间指标值时间序列的维度参差不齐；2)每位患者到医院检查的时间点不同，使得两条时间序列的检查天数可能不相同。因此，此节主要针对以上两个方面对时间序列进行同步化处理，完成聚类的准备工作。

4.1 不同时间序列的维度同步化

根据临床检验指标时间序列数据具有维度不同的具体特点，首先统计所有时间序列各维度区间的患者人数，然后利用式(1)计算得到同步化之后的时间序列维度。

$$Length=\frac{\sum_{i=1}^{n}T_i\ \text{中位数}\times N_i}{\sum_{i=1}^{n}N_i} \tag{1}$$

其中，$Length$ 表示规则化后每位患者指标值时间序列的维

度，T_i 中位数表示在 L_i 区间内所有患者指标序列长度值升序排列的中位数，N_i 表示 L_i 区间内患者序列的个数。

以本次实验数据为例，根据该数据特点将序列长度分为5个区间，分别以 L_i 表示。其中，$L_1=(5,10]$，此长度区间包含的患者数以 N_i 表示，经计算得出 $N_1=79$，然后依次统计计算，所得结果如表3所列。

表3 分阶段患者数表示示例

L	(5,10]	(10,15]	(15,20]	(20,25]	(25,30]
N	79	157	89	132	67

若以此数据为例，将区间 L_1 内的患者维度进行升序排列，然后可以得到该区间内的维度 T_1 中位数等于8；以此类推，可以得到每一个区间内的 T；最后根据式(1)，可以求得所有时间序列最佳的统一维度 $Length=12$。

4.2 不同时间序列时间点同步化

不同临床检验指标时间序列的时间点同步化处理旨在将不同时间序列的时间点置为相同的数值。经过上述的维度同步化之后，已得到时间序列的共同维度 *Length*。根据此维度值及所有时间序列的最大维度，计算出所有时间序列的统一时间点 $T=\{t_1,t_2,t_3,\cdots,t_{Length}\}$。然后利用 GetV 算法，计算得到每个 T_i 时间点所对应的指标值 V_i。

算法1 GetV 算法

输入：1. 临床检验指标值时间序列 TS[][]；//两维动态数组，其中第一维为指标值 v_i，第二维表示时间点 t_i

2. 统一时间点数组 T[Length]。

输出：同步化的指标值时间序列 Norm[][]。//两维动态数组，其中第一维为指标值 v_i，第二维表示时间点 t_i

```
//初始化动态数组 Norm[][];
1. for i=0 to T. length
2.      Norm[1][i]←T[i];
3. Endfor;
//计算同步后点值 v;
4. for i=0 to T. length
5. for j=0 to TS[1]. length
6.          If(j+1<TS[1]. length)
7. If(Norm[1][i]>TS[1][j]&&Norm [1][i]<TS[1][j+1])
8. Norm[0][i]←|(TS[0][j]-TS[0][j+1])/(TS[1][j]-TS[1][j]) * (Norm[1][i]-TS[1][j])+TS[0][j]|;
9. Break;
10.    Else if Norm[1][i]=TS[1][j]
11. Norm[0][i]←TS[0][j];
12. Else if
13. Norm[1][i]>TS[1][j];
14. Norm[0][i]←|(TS[0][j-1]-TS[0][j])/(TS[1][j-1]-TS[1][j-1]) * (Norm[1][i]-TS[1][j-1])+TS[0][j-1]|;
15. End for;
16. End for;
17. ReturnNorm[][].
```

对于图1和图2所示的两条指标时间序列A和B，经过时间点同步化，得到如图3所示的同步化结果。

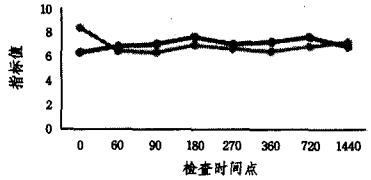

图3 A、B患者同步化后的指标值序列曲线

对于同步化之后的曲线，时间点是统一的，并且有各自的变化趋势。基于此就可以观察比较时间序列的差异性，并根据其变化趋势进行聚类研究。

5 非同步临床检验时间序列的聚类算法

本节给出经同步处理之后的临床检验指标时间序列聚类算法。由于患者个体差异，不同患者的疾病发展变化或多或少存在偏离，在临床检验指标上则体现为出现异常的时间序列点。为此，本文采用抗噪声能力较强的密度聚类方法，提出基于密度划分的临床检验指标时间序列聚类 LabTS-CLU 算法。

5.1 密度聚类算法概述

首先介绍基于密度聚类分析中所用到的几个基本定义。

定义1(邻域) 给定对象集合半径内的区域称为该对象的邻域。

定义2(核心对象) 如果给定对象邻域内的样本点数大于或等于 MinPts，则称该对象为核心对象。

定义3(直接密度可达) 给定一个对象集合 D，如果 p 在 q 的邻域内，且 q 是一个核心对象，则称对象 p 从对象 q 出发是直接密度可达的(directly density-reachable)。

定义4(密度可达) 对于样本集合 D，如果存在一个对象链 $p_1,p_2,p_3,\cdots,p_n,p_1=q$ 且 $p_n=p$，对象 p_{i+1} 是从 p_i 直接密度可达的，则对象 p 是从对象 q 密度可达的(density-reachable)。

定义5(密度相连) 如果存在对象 D，使对象 p 和 q 都是到对象 w 密度可达的，那么对象 p 到 q 密度相连的(density-connected)。

可以发现，密度可达是直接密度可达的传递闭包，并且这种关系是非对称的，只有核心对象之间相互密度可达。然而，密度相连是对称关系。dbscan 的目的是找到密度相连对象的最大集合。

算法2 dbscan 聚类算法

输入：1. ε——半径；

2. MinPts——给定点在邻域内成为核心对象的最小邻域点数；

3. D——数据集合。

输出：目标类簇集合。

方法：repeat

1. 判断输入点是否为核心对象；
2. 找出核心对象的邻域中的所有直接密度可达点；

until 所有输入点都判断完毕；

repeat

针对所有核心对象的邻域所有直接密度可达点，找到最大密度相连对象集合；

until 所有核心对象的邻域都遍历完毕；

5.2 LabTS-CLU 算法描述

LabTS-CLU 算法利用基于密度聚类的基本思路，即要求聚类空间中的一定区域内所包含对象(点或其他空间对象)的数目不小于某一给定阈值。给定一个临床检验指标时间序列集 $LabTS(L_1,\cdots,L_n)$，LabTS-CLU 算法将时间序列 L_i 视为多维空间中的一个对象点，在此基础上通过过滤低密度区域，将稠密度相对集中的对象点归为类 C_i，即将具有相同发展趋势特征的时间序列聚在同一个类，不同趋势特征的时间序列

聚在不同的类。

具体而言，LabTS-CLU 算法以前面经过同步处理后的时间序列集 LabTS 为基础，以 dbscan 算法为核心，完成时间序列的聚类工作。但是，我们在 LabTS-CLU 算法里引入了一个新的变量标准值 NoisePro，即聚类噪声点数的占有率，以此来满足用户对聚类结果的需求，并考察 LabTS-CLU 聚类算法的性能。

算法 3 LabTS-CLU 聚类算法

输入：1. E——针对序列集 LabTS 的密度半径；

2. PtsNum——基于上述序列集半径内的最小点数；

3. 临床检验指标时间序列集 LabTS(L_1，…，L_n)；

4. NoisePro—— 噪声点数的占有率。

输出：目标聚类簇。

方法：repeat

1. 基于序列集 LabTS 调用 dbscan 算法；

2. 调整参数 E，PtsNum；

3. 计算 NoisePro；

4. 返回步骤 1；

until NoisePro 达到余弦设定值；

与传统的基于划分聚类的凸形聚类簇不同，该算法可以发现任意形状的聚类簇。与传统的算法相比，它有如下优点：1)不需要输入要划分的聚类个数；2)聚类簇的形状没有偏倚；3)可以在需要时输入过滤噪声的参数。

6 实验对比

实验数据集来自某三甲医院，是经过隐私处理的病人的真实数据。经冗余数据处理、格式统一等过程，选取其中 10000 条作为实验数据。这 10000 条病人数据中，病人的维度从 5～34 不等，为了验证实验算法的差异，将数据维度统一规划为 4 个，即 5，10，20，30，并分别在这些维度的基础上考查算法效率及性能。本文均在同一实验平台进行研究，采用 eclipse 编程工具，使用 java 语言实现聚类算法。

对 k-means 和 LabTS-CLU 算法在不同数据量以及不同维度上进行了全面的效率和效果评价对比实验。对于 k-means算法，设计聚类簇数 $k=3$ 进行实验；对于 LabTS-CLU 算法，根据数据本身的特性，经过不断的测试，令半径 $\varepsilon=0.985$，最小点数集 $MinPts=10$。

首先对不同数据量下的两种算法的效果进行比对。主要针对算法的两方面进行研究：1)聚类后的效果，即数据聚类准确率；2)算法聚类的时间消耗，单位为 s。为了实验的准确性，首先对两种算法采用相同的维度，即令维度 $Length=12$，实验结果如表 4 所列。

表 4 与 k-means 聚类算法的性能比对

数据集	聚类算法	正确率(%)	耗时(s)
1000	k-means	75.67	15.69
	LabTS-CLU	89.45	11.42
10000	k-means	71.32	195.63
	LabTS-CLU	86.43	141.54

由表 4 可知，LabTS-CLU 的正确率明显高于 k-means 算法的，这也正是由算法真正的内在机制所决定：k-means 算法属于硬性聚类的算法，很机械地将数据聚为事先决定的类簇数，这必然会导致部分数据聚类错误；而 LabTS-CLU 算法是基于密度划分，属于较柔和的一种聚类算法，它会根据数据本身的特点结合用户输入的参数，将数据聚为事先不确定的几个类。虽然在此过程中，LabTS-CLU 算法形成的聚类会产生部分噪音点，但是从实验结果来看其效果还是优于 k-means 算法。

另外，随着数据量的增大，发现两种算法的正确率均有降低，这也在正常误差之内。数据量变大之后，势必会增加噪音点的个数，聚类中产生的错误点数也会随之增加，但是只要控制在误差范围之内，仍然可以说明算法的优越性。

对于表 4，单从时间来看可以得出两个结论：1)在相同数据量的基础上，LabTS-CLU 算法的时间消耗明显较小，至少可以缩减 15%～20%，这在效率上比 k-means 算法有相当的优越性；2)不同的数据量对算法的执行时间的影响较大。因此，若要最大程度地提高算法的执行效率，则需要考虑数据的分布式处理。

另外，在以上各参数设定的情况下，还通过实验从正确率以及耗时方面比对了 DBScan 聚类算法与 LabTS-CLU 算法，结果如表 5 所列。

表 5 与 DBScan 聚类算法的性能比对

数据集	聚类算法	正确率(%)	耗时(s)
1000	DBScan	84.53	10.25
	LabTS-CLU	89.45	11.42
10000	DBScan	80.47	138.49
	LabTS-CLU	86.43	141.54

由此实验可以看出，LabTS-CLU 算法与 DBScan 算法耗时相差不大，这也是由算法内部机制决定的。本文提出的 LabTS-CLU 算法是在 DBScan 算法的基础上进行改进的，因此算法的执行时间可以与原 DBScan 算法维持在同一数量级，保持了算法的时效性。

但是，实验结果显示 LabTS-CLU 算法在正确率方面远远优于 DBScan 算法，这也是我们对原算法进行改进之后的效果。由此可以知道，改进后的 LabTS-CLU 算法是有效的。

此外，还在不同维度上对 LabTS-CLU 算法进行了比较，可以明确地观察出数据的不同维度对 LabTS-CLU 算法的影响，效果如图 4 所示。

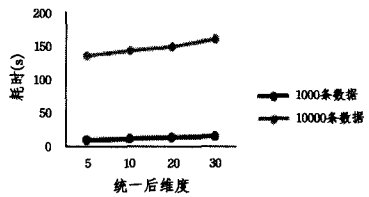

图 4 LabTS-CLU 算法在不同维度上的实验效率

由图 4 可以看出，在相同的数据量的基础上，随着维度的不断增加，算法的时间消耗是逐步增多的，但是增幅不大，即维度并不是算法耗时的主要影响因子；从纵向来看，相同维度下的数据量不同时，算法的时间消耗差异很大。由此可以断定，对于 LabTS-CLU 算法，数据集的大小才是算法耗时的主要影响因子。

结束语 本文针对临床数据的具体特点，研究了非同步临床检验指标时间序列的聚类问题，提出了一种非同步临床检验指标聚类 LabTS-CLU 算法。根据实验对比，本文的

（下转第 145 页）

Journal of Astronautics,2010,31(11):2584-2590(in Chinese)
焦健,张钦宇,李安国. 深空通信文件传输协议的交织技术研究[J]. 宇航学报,2010,31(11):2584-2590

[7] Mackay D J C. Fountain codes[J]. IEE Proceedings Communications,2005,152(6):1062-1068

[8] Byers J W,Luby M,Mitzenmacher M,et al. A digital fountain approach to reliable distribution of bulk data[C]// ACM SIGCOMM Computer Communication Review,1998,28(4):56-67

[9] Byers J W, Luby M, Mitzenmacher M. A digital fountain approach to asynchronous reliable multicast[J]. IEEE Journal on Selected Areas in Communications,2002,20(8):1528-1540

[10] Garrammone G,De C T,Matuz B,et al. Erasure codes for space communications:Recent findings and new challenges[C]// Advanced Satellite Multimedia Systems Conference (ASMS) and 12th Signal Processing for Space Communications Workshop (SPSC). 2012:29-35

[11] Zhu Kai-yan,Wang Hong-yu,Sun Wen-zhu,et al. A Distributed Fountain Code for Cooperative Communications[J]. Acta Electronica Sinica,2014,42(7):1249-1255(in Chinese)
祝开艳,王洪玉,孙文珠,等. 一种分布式喷泉码在协作通信中的应用[J]. 电子学报,2014,42(7):1249-1255

[12] Luby M. LT codes[C]// Proc of 43rd Annual IEEE Symposium on Foundations of Computer Society. 2002:271-271

[13] Cao Rui, Yang Liu-qing. Decomposed LT codes for cooperative relay communications[J]. IEEE Journal on Selected Areas in Communications,2012,30(2):407-414

[14] Sorensen J H,Koike-Akino T,Orlik P. Rateless feedback code [C]// The 2012 IEEE International Symposium on Information Theory. 2012:1767-1771

[15] Yu Jia-qi, Zhong Jie, Zhao Min-jian, et al. Novel LT coding scheme with limited feedback in broadcast systems[C]// The 2012 International Conference on Wireless Communications and Signal Processing(WCSP). 2012:1-5

[16] Hagedorn A, Agarwal S, Starobinski D, et al. Rateless coding with feedback[C]// International Conference on Computer Communications. 2009:1791-1799

[17] Xiao Jun-yuan,Li Ping. Doped accumulate LT codes[C]// IEEE International Symposium on Information Theory. 2007: 2001-2005

[18] Ying Ting, Xie Lei, Chen Hui-fang. A novel rateless coding scheme with gradually incremental degree under feedback[C]// IEEE Consumer Communications and Networking Conference (CCNC). 2014:575-580

[19] Kokalj-Filipovic S, Spasojevic P, Soljanin E, et al. Arq with doped fountain decoding[C]// International Symposium on Spread Spectrum Techniques and Applications. 2008:780-784

[20] Shokrollahi A. Raptor codes[J]. IEEE Transactions on Information Theory,2006,52(6):2551-2567

[21] Gu Shu-shi, Jiao Jian, Yang Zhi-hua, et al. Relay cooperation fountain codes for deep space communications[J]. Systems Engineering and Electronics,2012,34(8):1696-1701(in Chinese)
顾术实,焦健,杨志华,等. 面向深空通信的中继协作喷泉码设计[J]. 系统工程与电子技术,2012,34(8):1696-1701

(上接第 123 页)

LabTS-CLU算法明显优于基于 k-means 的聚类算法。另一方面,实验结果还表明,数据量是影响 LabTS-CLU 算法效率的主要因素,维度对其影响不大。

参 考 文 献

[1] Fravolini M L,Cascianelli S. Pier Giorgio Fabietti:A Learning Strategy for the Autonomous Control of Type 1 Diabetes[J]. Applied Artificial Intelligence,2015,29(6):531-562

[2] Goodwin G C,Medioli A M,Carrasco D S,et al. YongjiFu:A fundamental control limitation for linear positive systems with application to Type 1 diabetes treatment[J]. Automatica,2015,55:73-77

[3] Kellner D,Klappstein J,Dietmayer K. Grid-based DBSCAN for clustering extended objects in radar data[C]// 2012 IEEE Conference on Intelligent Vehicles Symposium. 2012:365-370

[4] Aghabozorgi S R, WahTeh Y. Clustering of large time series datasets[J]. Intelligent Data Analysis,2014,18(5):793-817

[5] Li Bin, Tan Li-xiang, Zhang Jing-song. Time series symbolic methods facing data mining[J]. Journal of Circuit and Systems,2000,5(2):9-14(in Chinese)
李斌,谭立湘,章劲松. 面向数据挖掘的时间序列符号化方法研究[J]. 电路与系统学报,2000,5(2):9-14

[6] Li Ai-guo,Qin Zheng. On-line segmentation of time-series data [J]. Journal of Software,2004,15(11):1671-1679(in Chinese)
李爱国,覃征. 在线分割时间序列[J]. 数据软件学报,2004,15(11):1671-1679

[7] Keogh E, Kasetty S. On the need for time series data mining benchmarks:A survey and empirical demonstration[J]. Data Mining and Knowledge Discovery,2003,7(4):349-371

[8] Tewari G, Snyder J, Sander P V. Signal-speciallized parameterization for piecewise linear reconstruction[C]// Proceedings of the Eurographics Symposium on Ueometry Processing. New York,USA,2004:55-64

[9] Tseng Y J,Ping Xiao-ou,Liang J D,et al. FeipeiLai:Multiple-Time-Series Clinical Data Processing for Classification With Merging Algorithm and Statistical Measures[J]. IEEE Journal of Biomedical and Health Informatics,2015,19(3):1036-1043

[10] Sitaram R, Zhang H, Guan C, et al. Temporal classification of multi-channel near-infrared spectroscopy signals of motor imagery for developing abrain-computer interface[J]. NeuroImage,2007,34(4):1416-1427

[11] Yin Z,Zhang J. Identification of temporal variations in mental workload using locally-linear embedding based EEG feature reduction andsupport vector machine based clustering and classification techniques[J]. Comput. Methods Programs Biomed.,2014,115(3):119-134