文章编号:1000-5862(2013)01-0056-04

一种有效的多元时间序列相似性度量算法分析

郭小芳[1],李　锋[2],刘庆华[1]
(1. 江苏科技大学计算机科学与工程学院,江苏 镇江 212003;
2. 江苏科技大学电子信息学院,江苏 镇江 212003)

摘要:为验证 Eros 距离对 MTS 数据集相似性度量的有效性,针对不同 MTS 数据集进行了相似性搜索实验研究. 结果表明:相对于其他的传统多元时间序列相似性度量,基于 Eros 距离的相似性度量方法比传统的方法在查全率-查准率上具有更大的优越性.

关键词:多元时间序列;相似性度量;欧几里德距离;扩展 Frobenius 范数
中图分类号:TP 391　　**文献标志码**:A

0　引言

相似性度量是各种多元时间序列相似性查询的基础,它直接影响着查询的通用性和完备性,同时也影响到序列的索引方法. 时间序列相似性度量主要通过距离度量方法来完成,如欧几里得距离(euclidean distance,ED)[1],动态时间弯曲距离(dynamic time warping,DTW)[2],加权奇异值分解(weighted sum singular value decomposition,W_{SSVD})[3] 和相似因子主元分析(PCA similarity factor,S_{PCA})[4]. 这几种方法都有各自的优缺点,如传统的欧几里德距离对时间序列的线性漂移和时间轴弯曲就不支持,它能很好地体现序列间的整体关系,但序列的局部变化(例如起伏变化)就不支持,这就有可能导致查询结果不正确. 而应用最多的动态弯曲距离计算效率较低,缺乏海量的索引技术.

实际的时间序列数据集中包含了多种因素影响的结果,各种因素之间的关系错综复杂,但其主要因素之间的关系可能是比较简单的. 如何快速有效地对数据进行分析,获得隐藏在所研究数据背后的变量关系,一直是个比较困难的问题. 采用基于主元范数的方法来分析 2 个多元时间序列的相似性,其主要步骤如下:(ⅰ)估算2个MTS元的协方差矩阵[5];(ⅱ)计算其特征值和特征向量;(ⅲ)通过在 MTS 数据集中获得的特征值来测量相应的每个 MTS 元的相似性. 验证实验中将 Eros 与不同的相似性度量方法作比较,结果表明 Eros 在查全率和查准率上要好于欧氏距离和动态时间弯曲距离.

1　相似性度量

对于实际中的多元时间序列 $X = \langle x_1, x_2, \cdots, x_j, \cdots, x_m \rangle$ 和 $Y = \langle y_1, y_2, \cdots, y_j, \cdots, y_m \rangle$,设定阈值 ε,若有 $D(X,Y) \leqslant \varepsilon$,则称时间序列 X 和 Y 相似,记作 $Sim(X,Y)$,其中 $D(X,Y)$ 为时间序列 X 和 Y 之间的距离函数,也称相似性度量函数.

1.1　欧氏距离相似度量

如用距离来衡量 2 个序列的相似程度,若 2 个序列对象之间距离越小,则其相似度越大,反之亦然. 时间序列相似性度量的 2 种经典方法分别是欧几里德距离和动态时间弯曲距离.

欧氏距离将时间序列(长度为 n)看作是 n 维欧氏空间中的一个点,其中两点 $X = \{x_1, x_2, \cdots, x_n\}$,$Y = \{y_1, y_2, \cdots, y_n\}$ 间的欧氏距离为

$$D(X,Y) = \sqrt{\sum_{i=1}^{n} | x_i - y_i |^2}. \quad (1)$$

欧氏距离有界,对坐标平移和旋转保持不变,计算简单,但这种相似性距离度量只能适应变化范围小的特定领域,对于振幅数值不同的序列,有时它们的形态变化相似,但之间的距离却可能相差很大,相反有时相似性距离很小的序列,它们之间的形态可

收稿日期:2012-08-12
基金项目:国家自然科学基金(51008143)资助项目.
作者简介:郭小芳(1974-),女,陕西商洛人,讲师,硕士,主要从事时间序列数据挖掘方面的研究.

能有很大差别. 这是其自身噪声与波动性的特点造成的,相似序列会呈现如波动干扰、平移偏移、振幅伸缩、线性漂移和不连续等多种变形现象.

1.2　动态弯曲距离相似度量

在 DTW 距离计算中,弯曲路径的计算必须满足如下特性: (ⅰ) 端点对齐: 弯曲路径起始于两序列的起点,结束于两序列的终点,即 $w_1 = (1,1)$, $w_k = (m,n)$; (ⅱ) 邻点连续: 弯曲路径上的任意 2 个相连点,在矩阵中是相邻的对应元素(可以是对角线相邻或者边界相邻),即对于给定 $w_k = (a,b)$,有 $w_k = (a',b')$,其中: $a - a' \leqslant 1$ 且 $b - b' \leqslant 1$; (ⅲ) 时间递增: 弯曲路径上的元素点是随时间增加而单调发展的,即对于给定 $w_k = (a,b)$,有 $w_k = (a',b')$,其中 $a - a' \geqslant 0$ 且 $b - b' \geqslant 0$. 图 1 中给出一条时间弯曲路径 W 与实际序列 Q 与 C 之间的映射关系.

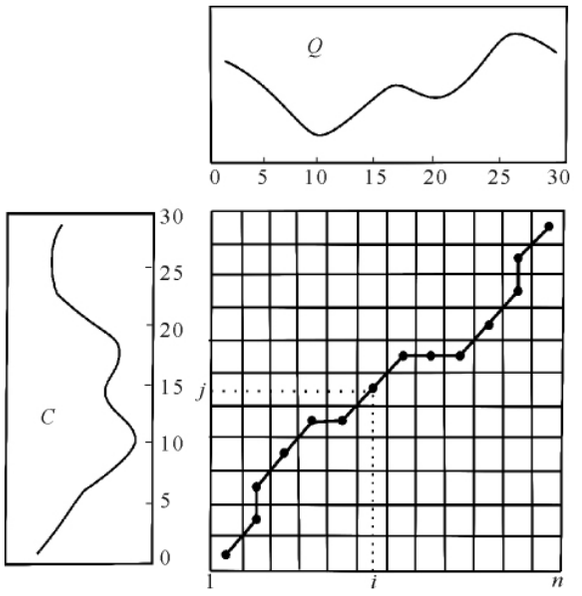

图 1　动态弯曲路径

动态时间弯曲距离根据两序列间各个数据点间的欧氏距离 $d(x_i,y_i)$ 所构成的矩阵,找出一条最短的有界、连续、单调的线段所构成的弯曲路径,该路径长度就是动态时间弯曲距离[7] 为

$$DTW(\boldsymbol{X},\boldsymbol{Y}) = \min\left\{\sqrt{\sum_{i=1}^{k} d(x_i,y_i)}/K\right\}, \quad (2)$$

其中 k 为较长序列的长度. DTW 支持不等长序列匹配,具有较高的识别和匹配精度,但计算的时间复杂度高,且不满足三角不等式,对检索的完备性是无法保证的,也不能很好地过滤不相似的序列[8]. 所以降低该算法的时间复杂度是该方法应用的关键所在.

2　扩展的 Euclid 范数距离

2.1　扩展 F 范数(Eros)

对于 MTS 项 $\boldsymbol{A}_{m_A\times n}$ 和 $\boldsymbol{B}_{m_B\times n}$,其右特征向量矩阵为 $\boldsymbol{V}_A = [\boldsymbol{a}_1,\cdots,\boldsymbol{a}_n]$ 和 $\boldsymbol{V}_B = [\boldsymbol{b}_1,\cdots,\boldsymbol{b}_n]$,将多变量时间序列进行主成分分析,并取各主成分的贡献率作为权重向量 $\boldsymbol{w}$,则 $\boldsymbol{A}_{m_A\times n}$ 和 $\boldsymbol{B}_{m_B\times n}$ 的扩展 F 范数(Eros) 定义为

$$Eros(\boldsymbol{A},\boldsymbol{B},\boldsymbol{w}) = \sum_{i=1}^{n} w_i \mid \langle a_i,b_j\rangle \mid = \sum_{i=1}^{n} w_i \mid \langle\cos\theta_i\rangle \mid, \quad (3)$$

权重向量 $\boldsymbol{w}$ 为一个对角半正定矩阵,满足 $w_{ii} \geqslant 0$ 且 $\sum_{i=1}^{n} w_i = 1$, $\langle \boldsymbol{a}_i,\boldsymbol{b}_i\rangle$ 是 $\boldsymbol{a}_i$ 和 $\boldsymbol{b}_i$ 的内积,θ_i 是 $\boldsymbol{a}_i$ 和 $\boldsymbol{b}_i$ 之间的夹角. Eros 大小从 0 到 1,可以反映 $\boldsymbol{A}_{m_A\times n}$ 和 $\boldsymbol{B}_{m_B\times n}$ 的相似程度,其中 1 为最相似.

2.2　Eros 距离的上下界

相似性查询的关键问题是相似性度量 $D(\boldsymbol{X},\boldsymbol{Y})$ 的选取[9],要求 $D(\boldsymbol{X},\boldsymbol{Y})$ 满足的条件为: (ⅰ) 非负性: $D(\boldsymbol{X},\boldsymbol{Y}) \geqslant 0$; 当且仅当 $\boldsymbol{X} = \boldsymbol{Y}$ 时,$D(\boldsymbol{X},\boldsymbol{Y}) = 0$; (ⅱ) 对称性: $D(\boldsymbol{X},\boldsymbol{Y}) = D(\boldsymbol{Y},\boldsymbol{X})$; (ⅲ) 三角不等式: $D(\boldsymbol{X},\boldsymbol{Y}) \leqslant D(\boldsymbol{X},\boldsymbol{Z}) + D(\boldsymbol{Z},\boldsymbol{Y})$.

考虑到 Eros 不满足距离三角形不等式,引入 $\boldsymbol{A}$ 和 $\boldsymbol{B}$ 之间的 Eros 距离为

$$D_{\text{Eros}}(\boldsymbol{A},\boldsymbol{B},\boldsymbol{w}) = \sqrt{2 - 2\sum_{i=1}^{n} w_i \left| \langle \boldsymbol{a}_i,\boldsymbol{b}_i\rangle \right|} = \sqrt{2 - 2\sum_{i=1}^{n} w_i \left| \sum_{j=1}^{n} a_{ij}b_{ji} \right|}. \quad (4)$$

可见,当 $\boldsymbol{A}$ 和 $\boldsymbol{B}$ 的相似度大于 $\boldsymbol{B}$ 和 $\boldsymbol{C}$ 的相似度,则有 $\text{Eros}(\boldsymbol{A},\boldsymbol{B},\boldsymbol{w}) > \text{Eros}(\boldsymbol{B},\boldsymbol{C},\boldsymbol{w})$,从而

$$D_{\text{Eros}}(\boldsymbol{A},\boldsymbol{B},\boldsymbol{w}) = \sqrt{2 - 2\sum_{i=1}^{n} w_i \left| \langle \boldsymbol{a}_i,\boldsymbol{b}_i\rangle \right|} < \sqrt{2 - 2\sum_{i=1}^{n} w_i \left| \langle \boldsymbol{b}_i,\boldsymbol{c}_i\rangle \right|} = D_{Eros}(\boldsymbol{B},\boldsymbol{C},\boldsymbol{w}), \quad (5)$$

即 $D_{\text{Eros}}(\boldsymbol{A},\boldsymbol{B})$ 小于 $D_{\text{Eros}}(\boldsymbol{B},\boldsymbol{C})$,可见 D_{Eros} 也可用于相似性度量. 为计算 2 个 MTS 项之间的相似度,将相关项的特征值作为权因子,比较 MTS 项的相应主元之间的相似度[10].

由于加权欧几里德距离满足

$$\sum_{i=1}^{n} w_i \sum_{j=1}^{n} \left| a_{ij}b_{ij} \right| \geqslant \sum_{i=1}^{n} w_i \left| \sum_{j=1}^{n} a_{ij}b_{ij} \right| \geqslant \sum_{i=1}^{n} w_i \sum_{j=1}^{n} a_{ij}b_{ij}, \quad (6)$$

所以

$$\sqrt{2-2\sum_{i=1}^{n}w_i\sum_{j=1}^{n}a_{ij}b_{ij}}\geqslant D_{\mathrm{Eros}}(\boldsymbol{A},\boldsymbol{B},\boldsymbol{w})\geqslant\sqrt{2-2\sum_{i=1}^{n}w_i\sum_{j=1}^{n}\left|a_{ij}b_{ij}\right|},\tag{7}$$

即 $D_{\mathrm{Eros}}(\boldsymbol{A},\boldsymbol{B},\boldsymbol{w})$ 存在上下界 $D_{\max}$,$D_{\min}$:

$$D_{\max}(\boldsymbol{A},\boldsymbol{B},\boldsymbol{w})=\sqrt{2-2\sum_{i=1}^{n}w_i\sum_{j=1}^{n}(a_{ij}-b_{ij})^2}=\sqrt{2-2\sum_{i=1}^{n}w_i\sum_{j=1}^{n}a_{ij}\times b_{ij}},\tag{8}$$

$$D_{\min}(\boldsymbol{A},\boldsymbol{B},\boldsymbol{w})=\sqrt{2-2\sum_{i=1}^{n}w_i\sum_{j=1}^{n}(|a_{ij}|-|b_{ij}|)^2}=\sqrt{2-2\sum_{i=1}^{n}w_i\sum_{j=1}^{n}\left|a_{ij}\times b_{ij}\right|},\tag{9}$$

因此,可通过计算公式 D_{Eros} 的上下界距离快速地过滤掉不满足相似要求的时间序列.

3 算法分析与数据实验

3.1 3 种相似度的比较

设 MTS 变量项数为 p,观测数为 m(通常 $p\ll m$),则根据(4)式和(6)式可知,使用 Eros 距离的运算量为 $O(mp^2+p^3)$,采用 DTW 距离的运算量为 $O(m^2p)$,采用 DTW 距离的运算量为 $O(mp)$. 若同时采用主成分分析的方法对数据进行降维处理,则计算 Eros 距离的计算复杂度将进一步降低. 尽管 DTW 支持时间伸缩,但当 $p\ll m$ 时其计算的代价高,且不满足三角不等式,不能直接用于大型时间序列数据库查询,对于有些索引方法还会造成查询结果遗漏.

为计算 2 个 MTS 项之间的相似度,Eros 将相关特征值作为权因子,比较 MTS 项的相应主元之间的相似度.

3.2 实验算法

MTS 经过主成分分析后,取各主成分的贡献率作为权重向量 w,采用修正留一 kNN[11] 对多元时间序列 MTS 进行查询.

算法 1 对 MTS 各项的协方差矩阵进行 SVD,得到本征值和右特征向量.

要求: 数据集中所有 MTS 项目 M.

for i = 1 to N do

$C\leftarrow M[i]$ 的协方差矩阵;

$[\boldsymbol{B},\boldsymbol{D},\boldsymbol{E}]\leftarrow SVD(\boldsymbol{C})$;

$s_i\leftarrow D$ 中的本征值;

将 E 存为第 i 个本征向量矩阵;

end for.

算法 2 计算权向量 $\boldsymbol{w}$.

要求: $\boldsymbol{S}_{n\times N}$ 矩阵,N 是 MTS 项目数,n 是变量数. $\boldsymbol{S}$ 中的列矢量 s_i 代表数据集中第 i 个 MTS 项目的所有本征值.

for i = 1 to N do

$s_i\leftarrow s_i\Big/\sum_{j=1}^{n}s_{ij}$

end for

for j = 1 to n do

$w_i\leftarrow f(\boldsymbol{s}_{\cdot,j})$;

end for

for i = 1 to n do

$w_i\leftarrow w_i\Big/\sum_{j=1}^{n}w_j$;

end for.

算法 3 该算法的查全率和查准率: 改进留一 kNN 搜索.

要求: MTS 项数 N,k,相应项的最大值 max r;

for i = 1 to 10 do

$precision[i]\leftarrow 0$;

end for

for i = 1 to N do

$Q\leftarrow M[i]$;

$k\leftarrow 1$,

$r\leftarrow 1$;

repeat

对 Q 执行 kNN 查询;

$c\leftarrow$ 被检索的 k 项中与 Q 相同标记项的数目;

if c = r then

$precision[r]\leftarrow precision[r]+c/k$;

$r\leftarrow r+1$;

end if

$k\leftarrow k+1$;

until $r\geqslant$ max r;

end for.

3.3 实验结果

对 MTS 作相似性查询,分别用 ED、D_{Eros} 以及 DTW 作为相似度量. 用 UCI 数据集[12] 中的脑电图(EEG)数据集作为测试数据集 Data_1,包含 122 个检测实例,每组记录的 4 个特征(变量数);日本元音字母数据集作为测试数据集 Data_2,包含 640 个检测实例,每组记 12 个变量.

表 1 给出了 D_{Eros}、ED 和 DTW 距离计算 2 个 MTS 相似性所用时间(elapsed time). 从表 1 可见,以消

耗时间衡量,对于不同的数据集,Eros 所用时间较少,Eros 的表现超过其他的相似性度量方法,这是因为 Eros 相似性度量方法对 MTS 项采取降维表示,时间复杂度低,其相对于欧氏距离和动态时间弯曲消耗的时间较少.

表 1　各种相似度所用时间/($\times 10^{-4}$S)

	$Data_1$	$Data_2$
D_{Eros}	2.71	0.453
ED	37.600	6.050
DTW	28.800	93.300

针对数据集 1,图 2 给出了 3 种相似度方法的查全率和查准率. 从图 2 可以看到,当查全率接近于 1 时,图 1 中 Eros 的查准率超过 *ED* 的 25%,超过 DTW 的 60%; 对于数据集 2,当查全率等于 0.1 时,Eros查准率至少优于 ED 和 DTW 的 6.5%,当查全率(recall) 等于 1 时,Eros 分别超过 ED 的 33%,超过 DTW 的 64%. 可见 D_{Eros}距离对不同属性的数据集都有较好的效果.

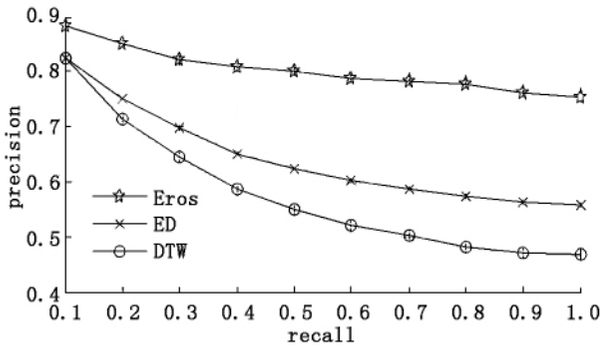

图 2　数据集 1 查全率查准率图

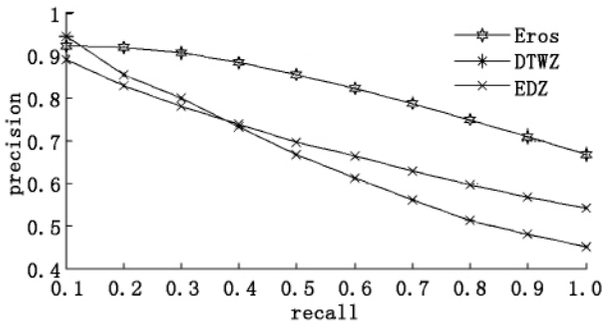

图 3　数据集 2 查全率查准率图

可见在两类 MTS 数据集上,Eros 相似度的查全率查准率高于其他两种方法,当对 2 个 MTS 项进行比较时,Eros 可以被用于实现高效率检索. 相对于其他的传统 MTS 相似性度量方法,这种 Eros 的多元时间序列相似性度量方法在查全率和查准率上具有更大的优越性. 因此验证了所提出的相似度测量方法的有效性.

4　结论

利用矩阵加权范数构造 MTS 主元之间范数距离(D_{Eros}) 用于 MTS 项的相似性度量. 分别采用了 ED、DTW 和 D_{Eros}距离针对两种数据进行了相似性查询实验. 实验结果表明基于范数距离(D_{Eros}) 的相似性查询方法的查全率和查准率明显超过基于 ED、DTW 的查询方法,且 D_{Eros}方法所用的时间最少.

5　参考文献

[1] Yang Kiyong, Shahabi C. An efficient *k* nearest neighbor search for multivariate time series [J]. Information Computation, 2007, 205(1): 65-98.

[2] 周大镯,吴晓丽,闫红灿. 一种高效的多变量时间序列相似查询算法 [J]. 计算机应用,2008,28(10): 2541-2543.

[3] 刘懿,鲍德沛,杨泽红,等. 新型时间序列相似性度量方法研究 [J]. 计算机应用研究,2007,24(5): 112-114.

[4] 陈胜利,李俊奎,刘小东. 基于提前终止的加速时间序列弯曲算法 [J]. 计算机应用,2010,30(4): 1068-1071.

[5] 曲文龙,张德政,杨炳儒. 基于小波和动态时间弯曲的时间序列相似匹配 [J]. 北京科技大学学报,2006,28(4): 396-402.

[6] Keogh E J, Ratanamahatana C A. Exact indexing of dynamic time warping [J]. Knowledge and Information Systems, 2004, 7(3): 154-158.

[7] Naseimento M, CarvalhoA. Spectral methods for clustering-a survey [J]. European Journal of Operational Researeh, 2011, 211(2): 221-231.

[8] Birant D, Kut A. ST-DBSCAN: an algorithm for clustering spatial-temporal data [J]. Data Knowledge Engineering, 2007, 60(1): 208-221.

[9] Chen Lei, Raymond Ng. On the marriage of Lp-norms and edit distance [EB/OL]. [2012-09-07]. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.1.7443&rep=rep1&type=pdf.

[10] Chen Lei, Tamer M Ö, Oria V. Robust and fast similarity search for moving object trajectories [EB/OL]. [2012-09-07]. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.94.8191&rep=rep1&type=pdf.

[11] 党育民. 序列模式挖掘算法研究 [J]. 江西师范大学学报: 自然科学版,2009,33(5): 604-607.

(下转第 73 页)

The Synthesis, Characterization and Spectroscopic Properties for Binary Rare Earth Complexes with 2-Hydroxy-6-Methylnicotinic Acid

ZHANG Yong[1], LIAO Li-ling[2], LIU Yu-bo[1], LI Cun-xiong[1,3*]

(1. Key Laboratory Information System of Mountains Area and Protection of Ecological Environment of Guizhou Province, Guizhou Normal University, Guiyang Guizhou 550001, China; 2. College of Chemistry and Materials Sciences, Guizhou Normal University, Guiyang Guizhou 550001, China; 3. College of Chemistry and Life Sciences, Guizhou Normal College, Guiyang Guizhou 5500183, China)

Abstract: Four new binary complexes of rare earths with 2-hydroxy-6-methylnicotinic acid(HA) were synthesized and characterized by FTIR spectra and elemental analysis, and carboxyl of the ligand chelated with rare earth ions in chelating bidentate mode. The composition of the complexes was confirmed to be $LnA_3 \cdot 3H_2O$(Ln = La, Eu, Gd, Tb). The fluorescence properties of Eu^{3+} and Tb^{3+} complexes were specially studied, two complexes both had good fluorescence propertie: $\eta(^5D_0 \rightarrow {}^7F_2 / {}^5D_0 \rightarrow {}^7F_1)$ of $EuA_3 \cdot 3H_2O$ is 5.9, $\eta(^5D_0 \rightarrow {}^7F_2 / {}^5D_0 \rightarrow {}^7F_1)$ of $TbA_3 \cdot 3H_2O$ is 1.9; 2-Hydroxy-6-methyl nicotine acid has a strong antenna effect on Eu^{3+} and Tb^{3+}; The measured fluorescence lifetime of $EuA_3 \cdot 3H_2O$ is 0.84 ms and $TbA_3 \cdot 3H_2O$ is 0.58 ms.

Key words: rare earth; 2-hydroxy-6-methylnicotinic acid; binary complex; spectroscopic properties

(责任编辑:刘显亮)

(上接第 59 页)

The Analysis for an Effective Algorithm of Similarity Measurement of Multivariate Time Series

GUO Xiao-fang[1], LI Feng[3], LIU Qing-hua[1]

(1. School of Computer Science and Engineering, Jiangsu University of Science and Technology, Zhenjiang Jiangsu 212003, China; 2. School of Electronics and Information, Jiangsu University of Science and Technology, Zhenjiang Jiangsu 212003, China)

Abstract: In order to show the validity of Eros for similarity search on MTS datasets, several experiments were performed on different datasets. The experimental results show that the method of similarity measurement based on Eros distance has superiority in Recall-Precision as compared to the traditional similarity measurements for MTS datasets.

Key words: multivariate time series; similarity measurement; Euclidean distance; extended Frobenius Norm(Eros)

(责任编辑:冉小晓)