基于弯曲距离三支决策的时序相似性算法

徐健锋[1,2] 何宇凡[2] 张远健[1] 汤 涛[2]
(同济大学计算机科学与技术系 上海 201804)[1] (南昌大学软件学院 南昌 330047)[2]

摘 要 动态时间弯曲距离算法(DTW)是目前公认的最有效的时间序列相似性计算方法之一,但是较高的时间复杂度一直是其主要缺点。快速弯曲距离算法(FTW)能有效提高DTW的计算速度,但是该算法对不同粒度时间序列剪枝的行为是典型的二支决策,与人类处理不确定问题时普遍采用的三支判断不同。因此,通过将三支决策理论引入到DTW算法的优化工作中,建立了DTW三支决策模型;然后对DTW三支决策模型中的决策阈值α和β进行了基于误识别率的推导,并且给出了具体求解阈值α和β的模拟退火算法;最后基于上述理论提出了基于弯曲距离三支决策的时序相似性算法(3WD-DTW)。通过对比实验表明,与FTW算法相比,3WD-DTW算法在保持较快的计算速度的前提下明显提升了计算准确度,使其接近DTW的水平。

关键词 三支决策,动态时间弯曲,模拟退火,决策阈值
中图法分类号 TP181 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2017.09.007

Similarity Algorithm Based on Three Way Decision of Time Warping Distance

XU Jian-feng[1,2] HE Yu-fan[2] ZHANG Yuan-jian[1] TANG Tao[2]
(Department of Computer Science and Technology, Tongji University, Shanghai 201804, China)[1]
(School of Software, Nanchang University, Nanchang 330047, China)[2]

Abstract Dynamic time warping (DTW) is widely accepted as one of the most effective methods for the similarity measurement of time series, but suffers from high time complexity. Fast search method for dynamic time wrapping (FTW) is demonstrated to accelerate DTW. The core of pruning is however a typical two-way decision rather than three-way decision, which is different from actions taken with uncertain issues. By incorporating three-way decision, an optimized DTW model three-way decision DTW (3WD-DTW) is developed first. The decision thresholds α, β are derived by solving an optimization problem with the objective of minimizing error rate. A novel simulated annealing algorithm is thus proposed. Finally, similarity algorithm based on three way decision of time warping distance is presented. Experiments show that 3WD-DTW is comparable in computing complexity as compared to FTW. In terms of accuracy, 3WD-DTW outperforms FTW significantly and approximates to DTW.

Keywords Three-way decision, Dynamic time warping, Simulated annealing, Decision threshold

1 引言

时间序列数据挖掘是数据挖掘的一个重要领域,其应用可以涉及到国民生产和生活的各个行业。时间序列数据挖掘已有一定的研究历史,且随着物联网、大数据时代的到来,对时间序列数据挖掘的研究越来越受到学术界和企业界的重视。

时间序列相似性计算是时间序列数据挖掘的基础研究之一,常用的方法有欧氏距离法、夹角余弦法、动态弯曲距离法等[1-2]。其中,动态时间弯曲距离(DTW)算法[3]是目前公认的抗干扰、抗变形效果最好的时间序列相似性计算方法,但是其时间复杂度较高的缺点也十分突出。

多年来,学术界对DTW算法的优化研究一直没有停止,成果也很多。综合来讲,对DTW的优化方式主要可以分为二大类:提前终止方法[4]和多粒度分层递进方法[5]。其中,多粒度分层递进的典型方法是Sakurai等人在ACM PODS2003会议上提出的Fast Similarity Search under the Time Warping (FTW)[6]。其主要思想是由粗粒度到细粒度对被测量序列进行分层递进的DTW分析,对每个粒度下的不相似序列进行提前剪枝,以达到大幅提高计算速度的效果。但是FTW算法的核心思想是采用典型的二分类策略对DTW算法进行优化,与人们实际处理不确定问题的三支决策经验有所不同[7],虽然其计算速度与经典DTW算法相比有极大提高,但其准确度却大幅降低。当前,如何在上述研究的基础上进一步提升该算法的计算准确度,是一个重要的议题。

三支决策是由加拿大著名学者姚一豫教授提出的一种求解不确定问题的新理论,该理论与粗糙集理论中的集合正域、

到稿日期:2016-07-27 返修日期:2016-09-07 本文受国家自然科学基金:粒计算中的不确定性分析与研究(61273304),上海市中医药三年行动计划重点项目:中医目诊仪(临床诊疗设备)开发研究(ZY3-CCCX-3-6002)资助。
徐健锋(1973-),男,博士生,副教授,主要研究方向为粒计算、粗糙集、三支决策,E-mail:940jianfeng_x@tongji.edu.cn;何宇凡(1994-),男,硕士生,主要研究方向为机器学习;张远健(1990-),男,博士生,主要研究方向为粒计算、粗糙集、三支决策。

负域和边界域概念相对应，决策者有接收、拒绝和延迟 3 种决策；姚教授同时也给出了超出决策粗糙集模型的普适三支决策的语义解释[8]。通过近几年的研究和发展，三支决策被认为在信息不足或者获取足够信息的代价较高时能够兼顾决策代价与正确性，其思想和方法与人类解决不确定问题的思维方式相吻合，是在认知方面具有优势和效益[9]的决策思想。近年来，三支决策被广泛应用于聚类[10-11]、分类[12-13]、属性约简[14-15]、邮件过滤[16-17]、图形处理[18-19]等各个领域。

时间序列相似性计算的难点正是由于其相似度的不确定性导致的。本文的主要贡献在于，将三支决策理论引入到 DTW 算法优化研究中，建立了 DTW 三支决策模型；然后对模型中的决策阈值[20-21]提出了基于误识别率的计算方案，并且给出了一种具体的模拟退火算法；最后基于上述理论提出了一种有效的基于弯曲距离三支决策的时序相似性算法，简称 3WD-DTW。

本文第 2 节主要介绍时序弯曲距离算法的基本概念及三支决策的基本思想理论；第 3 节给出 DTW 三支决策模型和三支决策阈值的求解方法；第 4 节通过对比实验对 3WD-DTW 算法的性能进行分析；最后总结全文。

2　相关技术

2.1　时序弯曲距离算法

动态时间弯曲算法(DTW)[3]是一种抗干扰、抗形变的相似性计算方法。其基本原理是：假设长度为 m 的时间序列 $P=\{p_1,p_2,\cdots,p_m\}$，长度为 n 的时间序列 $Q=\{q_1,q_2,\cdots,q_n\}$，如图 1 所示，根据匹配两序列中各元素间的距离形成距离矩阵；然后通过动态规划法找到一条连续的路径，使得该路径上的距离累加和最小。

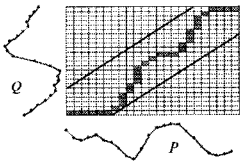

图 1　弯曲矩阵与弯曲路径

P 和 Q 间的动态时间弯曲距离记为 $DTW(P,Q)$，其匹配效果如图 2 所示。$DTW(P,Q)$的计算公式可表示如下：

$$DTW(P,Q)=d(m,n)$$

$$d(i,j)=(p_i-q_j)^2+\min\begin{cases}d(i,j-1)\\d(i-1,j)\\d(i-1,j-1)\end{cases}$$

$$d(0,0)=0,d(i,0)=d(0,j)=\infty$$

其中，$0\leqslant i\leqslant m,0\leqslant j\leqslant n$。

由上述公式可知，DTW 算法的时间复杂度为 $O(mn)$，此高复杂度严重制约了 DTW 的可使用性。

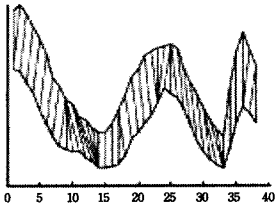

图 2　弯曲距离匹配效果图

动态时间弯曲下的快速相似性搜索算法(FTW)[6]的基本步骤如下。

步骤 1　对原始时间序列和目标时间序列同时进行粗粒度窗口划分，然后计算该粗粒度窗口下各时间序列与目标时间序列的 DTW 距离，根据计算的结果删除不相似程度较高的序列。

步骤 2　将余下的时间序列进行更细粒度的时序窗口划分，然后再计算该粒度窗口下各时间序列与目标时间序列的 DTW 距离，根据计算的结果删除不相似程度较高的序列。

步骤 3　如果当前时间序列的窗口粒度为最小，则算法结束，并且统计时间序列的相似性情况；否则进入步骤 2。

注：粒度的细分与相似程度阈值的设置有关。

该算法基于相似度二支决策，以分层递进的方式进行剪枝操作，以达到减少计算量、提升计算速度的目的。但是，显然该算法与人们实际处理不确定问题的三支决策经验有所不同。

2.2　三支决策理论简介

三支决策是由国际著名粒计算专家姚一豫[7]在粗糙集研究的基础上首先提出的一种新的求解不确定问题的理论。其初始目的是为粗糙集理论中的 3 个分类区域，即正域、负域和边界域，提供合理的决策语义解释。其基本思想如图 3 所示，将一个有限非空对象集合整体 $U=\{x_1,x_2,\cdots,x_n\}$分为 3 个独立的部分，简述为 $U\xrightarrow{f}\{P,B,N\}$，由正域 $P(U)$生成的正规则对应做出接收决策，由负域 $B(U)$生成的负规则对应做出拒绝决策，由边界域 $N(U)$生成的边界规则对应做出延期决策。

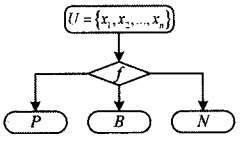

图 3　三支决策示意图

三支决策理论虽然是基于粗糙集研究提出的理论成果，但是并不应只局限于粗糙集合。当前已有学者进行了区间集[22]、模糊集[23]、阴影集[24]等非粗糙集三支决策模型的研究，基于此构建一个更普适且更具应用价值的理论是三支决策重要的研究方向。本文在 DTW 和 FTW 算法的研究中引入三支决策思想，建立 DTW 相似性三支决策模型，是将三支决策理论应用到非粗糙集模型框架的又一种探索。

3　时序相似性的三支决策

3.1　DTW 相似性三支决策

定义 1　设 $TIS^k=\{U,C^k,l,V,f\}$是时间窗口粒度大小为 k 的时序信息系统。

其中，$U=\{P_1,P_2,\cdots,P_n\}$表示时序数据对象的集合；C^k是时序数据的条件属性，粒度为 k。注：$k\in\{1,2,\cdots,h\}$代表时间序列数据窗口的大小。l 为时序数据的标签属性，即对象类别标签。V 为条件属性与类别标签的值域。f 表示 $U\to V$ 的信息映射，记作 $f:(C^k,l)\times P_i\to(C_i^k,l_i)$，其中 $P_i\in U$，

(C_i^k, l_i)的语义为第 i 个时间序列的条件部分取值与决策部分取值。$P_0=(C_0^k, l_0)$为被比较的目标样本，它具有与 $\forall P_i \in U$ 相同的数据结构。C_i^k 和 C_0^k 之间的时序弯曲距离记为 $DTW(C_i^k, C_0^k)$。

定义 2 DTW 相似性三支决策即依据 $DTW(C_i^k, C_0^k)$判定 P_i 和 P_0 之间是否相似的三支决策，可以定义如下：

1)若 $DTW(C_i^k, C_0^k) \leqslant \alpha$ 成立，即判定时间序列对象 P_i 与目标序列 P_0 相似，记为 $l_i \approx l_0$；

2)若 $DTW(C_i^k, C_0^k) \geqslant \beta$ 成立，即判定时间序列对象 P_i 与目标序列 P_0 不相似，记为 $l_i \not\approx l_0$；

3)若 $\alpha < DTW(C_i^k, C_0^k) < \beta$ 成立，即判定时间序列对象 P_i 与目标序列 P_0 相似性为不确定，记为 $l_i \simeq l_0$。

注：P_i 的标签客观上与 P_0 相同，则表示为 $l_i = l_0$；P_i 的标签客观上与 P_0 不同，则表示为 $l_i \neq l_0$。

3.2 基于误识率的 DTW 相似性三支决策阈值推导

对于时序数据的相似性计算而言，算法的识别率、误识率等指标通常是人们评价该算法优劣的依据。

定义 3 通过 DTW 算法判断为 $l_i \approx l_0$ 决策的误识率为 λ_{tF}，即算法预测为 $l_i \approx l_0$ 的对象集合中实际 $l_i \neq l_0$ 的对象所占的比例。λ_{tF} 的计算公式如下：

$$\lambda_{tF} = \Pr([P_i]_{l_i \neq l_0} / [P_i]_{DTW(C_i^k, C_0^k) \leqslant \alpha}) = \frac{|[P_i]_{DTW(C_i^k, C_0^k) \leqslant \alpha} \cap [P_i]_{l_i \neq l_0}|}{|[P_i]_{DTW(C_i^k, C_0^k) \leqslant \alpha}|}$$

定义 4 通过 DTW 算法判断为 $l_i \not\approx l_0$ 决策的误识率为 λ_{fT}，即算法预测为 $l_i \not\approx l_0$ 的对象集合中实际 $l_i = l_0$ 的对象所占的比例，λ_{fT} 的计算公式如下：

$$\lambda_{fT} = \Pr([P_i]_{l_i = l_0} / [P_i]_{DTW(C_i^k, C_0^k) \geqslant \beta}) = \frac{|[P_i]_{l_i = l_0} \cap [P_i]_{DTW(C_i^k, C_0^k) \geqslant \beta}|}{|[P_i]_{DTW(C_i^k, C_0^k) \geqslant \beta}|}$$

在机器学习算法的研究中，通常人们都希望算法的误识率越低越好，也经常设定可以接受的最低误识率。在 DTW 算法中当最大可接受误识率 λ_{tF} 和 λ_{fT} 分别设定为 λ_T 和 λ_F 时，误识率限定函数 $Er(\alpha)$和 $Er(\beta)$可表示为：

$$Er(\alpha) = \lambda_T - \lambda_{tF} = \lambda_T - \frac{|[P_i]_{DTW(C_i^k, C_0^k) \leqslant \alpha} \cap [P_i]_{l_i \neq l_0}|}{|[P_i]_{DTW(C_i^k, C_0^k) \leqslant \alpha}|}$$

$$Er(\beta) = \lambda_F - \lambda_{fT} = \lambda_F - \frac{|[P_i]_{l_i = l_0} \cap [P_i]_{DTW(C_i^k, C_0^k) \geqslant \beta}|}{|[P_i]_{DTW(C_i^k, C_0^k) \geqslant \beta}|}$$

显然，推导出阈值 α 和 β 的过程可以表示为约束最优化问题 $\arg\max_{\alpha}(Er(\alpha))$和 $\arg\max_{\beta}(Er(\beta))$的求解。

3.3 基于模拟退火法的 DTW 三支决策阈值求解算法

推导出阈值 α 和 β 的过程是一个典型的求解最优化的问题。模拟退火算法[20]是用来解决这一类问题的重要方法之一。其主要求解思路[19]是在解空间中随机选定一个初始当前解，基于该解从解空间中生成一个新解，计算新解和当前解的适应度函数之差，若该差值满足相应条件则接受新解作为新的当前解，否则以一定概率接受新解作为新的当前解。重复上述迭代过程，直到满足终止条件为止。

将推导决策阈值 α 和 β 的解空间解释为各粒度下不同时序之间的 DTW 距离集合。适应度函数设为 Er。冷却机制部分，初始温度借鉴文献[25]设为 $Tem = \frac{Er_{\min} - Er_{\max}}{\ln t_0}$，其中 t_0 表示初始的接受阈值，这里设 $t_0 = 0.5$，$Er_{\max}$ 和 $Er_{\min}$ 表示限定函数 Er 的极大值和极小值。温度 Tem 的变化遵循 $Tem = Tem \cdot r$，这里设置 $r = 0.9$。

构造基于模拟退火法的 DTW 三支决策阈值求解算法，具体步骤如下。

算法 1 基于模拟退火法的 DTW 三支决策阈值求解算法

输入：所有训练集与标准序列的动态弯曲距离，即 $S=\{s_i \mid s_i = DTW(C_i^k, C_0^k)(0 \leqslant i \leqslant n)\}$

输出：阈值 α, β

Step1 初始化：温度 Tem；解空间 S；迭代起点 $S_\alpha = S_0$ 和 $S_\beta = S_0$，分别表示求解阈值 α 和 β 的当前解；每个 Tem 值的迭代次数 n。

Step2 循环 Step3—Step5 n 次。

Step3 产生新解 S_α'，S_β'。

Step4 若 $Er(S_\alpha') > Er(S_\alpha)$或者 $e^{\frac{(Er(S_\alpha')-Er(S_\alpha))}{Tem}} > \rho$，则接受 S_α'作为新的当前解；若 $Er(S_\beta') > Er(S_\beta)$或者 $e^{\frac{(Er(S_\beta')-Er(S_\beta))}{Tem}} > \rho$，则接受 S_β'作为新的当前解。注：ρ 是一个介于 0 和 1 之间的随机数，作为相邻解被接收为当前解的概率。

Step5 如果连续若干个新解都没有被接受，则输出当前解作为最优解，得到阈值 α, β，结束程序。

Step6 $Tem = Tem \cdot r$，若 $Tem \leqslant Tem_{\min}$，则输出当前解作为最优解，得到阈值 α, β，结束程序；否则执行 Step2。

3.4 基于弯曲距离三支决策的时序相似性算法

本研究在 FTW 算法的基础上通过引入 DTW 相似性三支决策理论，将二支决策的剪枝方案转变为更加符合人类思维习惯的三支决策剪枝，提出算法 3WD-DTW。

其主要思想是由粗粒度到细粒度对被测量序列进行分层递进的 DTW 相似性三支决策，每一层粒度的划分体现在对时间序列在时间轴上的等间隔划分，其中第 h 层对应的间隔最大，粒度最粗；第 1 层对应的间隔最小，粒度最细。首先计算粗粒度下时间序列与目标时间序列的 DTW 相似性三支决策，对相似度小于或等于 α 的序列做出相似决策；对相似度大于或等于 β 的序列做出不相似决策；而对相似度在 α, β 之间的序列，则进入较细粒度下与目标时间序列做 DTW 相似性三支决策，以此类推直至最细粒度，重复上述三支决策步骤。

3WD-DTW 算法的具体步骤如下。

算法 2 3WD-DTW 算法

输入：k 级粒度的目标序列 P_0，k 级粒度的测试序列 P，k 级粒度下的三支决策阈值 α, β。注：$k \in \{1, 2, \cdots, h\}$，且 k 初始为 h。

输出：P 和 P_0 是否相似，若相似返回 $l_i \approx l_0$，否则返回 $l_i \not\approx l_0$。

Step1 计算粗粒度下 P 和 P_0 间的动态时间弯曲距离 $DTW(C_1^k, C_0^k)$。

Step2 若 k=1，则转至 Step4，否则转至 Step3。

Step3 若 $DTW(C_1^k, C_0^k) \leqslant \alpha$，则返回 $l_i \approx l_0$；若 $DTW(C_1^k, C_0^k) \geqslant \beta$，则返回 $l_i \not\approx l_0$；否则，P 和 P_0 细化为粒度 k，k=k−1，转 Step1。

Step4 若 $DTW(C_1^k, C_0^k) \leqslant \alpha$，则返回 $l_i \approx l_0$；否则返回 $l_i \not\approx l_0$。

本算法的时间复杂度为 $O(\frac{m}{k}\frac{n}{k})$，其中 m 和 n 分别表示目标序列 P_0 和测试序列 P 的时序数据长度。其与文献[6]中所述 FTW 算法的时间复杂度相同；但与 DTW 算法的时间复杂度 $O(mn)$相比，本算法的计算速度理论上可以提升 k^2 倍，故其能明显提高算法的计算速度。

4　实验与分析

4.1　实验数据描述

为了验证本文所提算法的有效性并保证实验的可信度，实验中所使用的数据集均来源于 UCR 数据库[26]。数据集的详细信息如表 1 所列，在实验中只考虑每个数据集的两个类，即属于标准类和不属于标准类。

表 1　数据集信息

SET NO	数据集	类别数	训练集大小/条	测试集大小/条	时序数据长度/个
SET 1	Worms	2	77	181	900
SET 2	Face (all)	2	560	1690	131
SET 3	Adiac	2	390	391	176
SET 4	Symbols	2	25	995	398
SET 5	Coffee	2	28	28	286
SET 6	OliveOil	2	30	30	570
SET 7	ShapesAll	2	600	600	512
SET 8	Lightning-7	2	70	73	319

本次实验采用 DTW、FTW、欧氏距离相似度比对算法（下文记作 ED）以及 3WD-DTW 算法分别对表 1 所述的 8 个数据集（SET1－SET8）进行相似性计算。由于文献[3,6]介绍的 DTW 算法和 FTW 算法的平均准确率分别为 85%和 75%，因此本实验采用两者的平均值 0.8 作为 λ_F 与 λ_T 的取值。实验所得 3WD-DTW 算法决策阈值 α，β 的结果如表 2 所列。表 2 中，k 表示对时序数据划分的粒度大小，α 和 β 表示 3WD-DTW 算法的决策阈值。结合表 2 中各个数据集中时间序列的数据大小信息和最后决策阈值的选取可知，时序数据的数据量越多，则阈值会相应越大，不同时间序列之间的差异程度越大；且对同一个数据集而言，粒度越粗，则阈值相应越小。这说明，时序数据的复杂程度对决策阈值的大小有正相关影响。

表 2　3WD-DTW 算法决策阈值表

	k=4		k=3		k=2		k=1
	α	β	α	β	α	β	α
SET 1	82.5	400	103	435	118	457	477
SET 2	4E−5	0.2	0.0	3.4	1.8	6.8	22.0
SET 3	3E−5	0.1	3E−3	0.08	0.01	0.11	0.3
SET 4	1.3	1.2	1.53	1.8	1.9	2.2	2.5
SET 5	0.1	0.5	0.2	0.7	0.4	1.4	2.3
SET 6	2E−4	3E−4	4E−4	8E−4	2E−3	2E−3	0.03
SET 7	11.9	11.0	13.4	11.7	14.5	12.4	13.2
SET 8	0.2	5.8	0.2	7.2	0.8	16.0	53.7

4.2　实验结果及分析

由于实验中采用不同数据集的数据进行实验，而每个数据集大小不同，因此本实验对计算速度进行归一化处理。具体将同一数据集的计算速度归一化为：

$$T_i = t_i / \sum_{j=1}^{3} t_j$$

其中，t_i 为第 i 个算法的消耗时间，$j=\{1,2,3\}$ 表示第 j 个算法，计算速度比 T_i 为归一化后的计算速度观测值。

实验结果如表 3 所列。可以看出，由于 ED 算法将样品对象的不同属性之间的差别等同看待，虽然其时间消耗上占优势，但其准确率却不理想。下划线部分代表给定数据集在某一评价指标下的最优效果。由表 3 中运行时间数据可以看出，3WD-DTW 算法很好地继承了 FTW 算法在计算速度上表现出来的优势。从大多数数据集中的实验结果可以看出，3WD-DTW 算法和 FTW 算法在计算速度上相差不大，且都大幅度优于 DTW 算法，其速度对比效果如图 4 所示。

表 3　实验结果对比表

	准确率/%				时间消耗/ms			
	3WD-DTW	DTW	ED	FTW	3WD-DTW	DTW	ED	FTW
SET 1	52.22	57.78	34.44	51.11	8924	15352	308	766
SET 2	76.97	89.22	54.74	60.57	795	3199	164	311
SET 3	99.49	99.49	63.59	96.92	235	1237	73	110
SET 4	96.88	97.99	61.97	96.18	618	15554	208	608
SET 5	62.96	66.67	40.74	59.26	87	312	19	32
SET 6	65.52	68.97	48.28	62.07	125	1121	45	63
SET 7	99.17	99.17	65.44	99.00	582	16470	323	562
SET 8	95.83	97.22	53.66	81.71	220	808	94	121

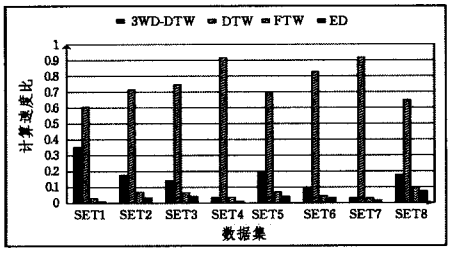

图 4　不同数据集下各算法的计算速度比对比图

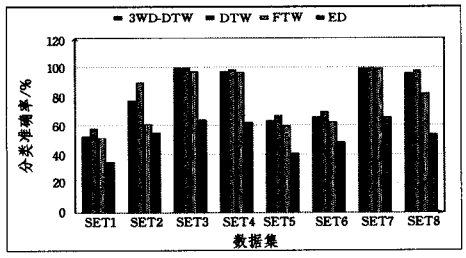

图 5　不同数据集下各算法的分类准确率对比图

由表 3 中算法准确率数据可以看出，3WD-DTW 算法的准确率明显优于 FTW 算法的准确率，且与 DTW 算法的准确率相差不大。例如在数据集 Adiac 中，3WD-DTW 的准确率接近 DTW 算法的准确率；且平均算法准确率比 FTW 算法提高了 5%。上述 4 种算法的分类准确率对比效果如图 5 所示。

对于 FTW 算法，虽然其平均计算速度比 DTW 算法的提高了 95%，但是其平均算法准确率却降低了 8%；而对于 3WD-DTW 算法，不但其平均计算速度比 DTW 算法的提高了 78.56%，而且平均算法准确率仅降低了 3%。也就是说，FTW 算法付出准确率降低 1%的代价仅提高了 11.87%的计

算速度，而 3WD-DTW 算法每付出准确率降低 1%的代价却能提高 26.19%的计算速度，显然 3WD-DTW 算法比 FTW 算法和 DTW 算法具有更好的性价比。3WD-DTW 算法的优越性在于采用了三支决策的思想，更加合理地解决了 DTW 计算中存在的不确定性现象，从而兼顾了计算速度与正确性。本算法明显优于只采用两支决策的 FTW 算法，更加证明了三支决策应用的普遍有效性。

结束语 本文将三支决策的思想应用到了经典的时间序列 DTW 算法的优化上，在建立时序信息系统的基础上提出了 DTW 相似性的三支决策模型；基于上述模型进行了基于算法误识率的 DTW 相似性三支决策阈值推导研究，并且具体给出了决策阈值模拟退火求解方法；最后基于上述理论研究设计了基于弯曲距离三支决策的时序相似性算法。通过对比实验表明，3WD-DTW 算法明显优于只采用两支决策的 FTW 算法，更加证明了三支决策应用的普遍有效性。

本研究建立的非粗糙集的 DTW 相似性三支决策模型是将三支决策理论应用在更广领域的有益探索。同时，DTW 相似性三支决策模型还有很多方面有待进一步完善，使其更加具有普适性。例如，本文只是讨论了基于误识率的 DTW 相似性三支决策阈值的推理关系，但是在许多实际应用中除了查准率还有查全率、延迟率等多种因素也是需要考虑的；λ_F 与 λ_T 的设定与阈值之间的关系也有待讨论；对于阈值的计算，本文虽然已提出了一种最优化问题求解方法，但是否还有其他更有效的求解方法也需要进一步的研究。

参 考 文 献

[1] GULLO F, PONTI G, TAGARELLI A, et al. A time series representation model for accurate and fast similarity detection[J]. Pattern Recognition, 2009, 42(11): 2998-3014.

[2] WANG Q, MEGALOOIKONOMOU V. A dimensionality reduction technique for efficient time series similarity analysis[J]. Information systems, 2008, 33(1): 115-132.

[3] KEOGH E, RATANAMAHATANA C A. Exact indexing of dynamic time warping[J]. Knowledge & Information Systems, 2010, 7(3): 358-386.

[4] LI H L, YANG L B. Extensions and relationships of some existing lower-bound functions for dynamic time warping[J]. Journal of Intelligent Information Systems, 2014, 43(1): 59-79.

[5] SUN L, YANG Y J, LIU W H. Trended DTW Based on Piecewise Linear Approximation for Time Series Mining[C]// 2011 IEEE 11th International Conference on Data Mining Workshops (ICDMW). IEEE, 2011: 877-884.

[6] SAKURAI Y, YOSHIKAWA M, FALOUTSOS C. FTW: fast similarity search under the time warping distance[C]// Proceedings of the Twenty-fourth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems. ACM, 2005: 326-337.

[7] YAO Y Y. An outline of a theory of three-way decisions[M]// Rough Sets and Current Trends in Computing. Springer Berlin Heidelberg, 2012: 1-17.

[8] YAO Y. Three-Way Decision: An Interpretation of Rules in Rough Set Theory[M] // Rough Sets and Knowledge Technology. Springer-Verlag, 2009: 642-649.

[9] YAO Y Y. Three-Way Decisions and Cognitive Computing[J]. Cognitive Computation, 2016, 8(4): 543-554.

[10] LI F, YE M, CHEN X. An extension to Rough c-means clustering based on decision-theoretic Rough Sets model[J]. International Journal of Approximate Reasoning, 2014, 55(1): 116-129.

[11] YU H, LIU Z G, WANG G Y. An automatic method to determine the number of clusters using decision-theoretic rough set [J]. International Journal of Approximate Reasoning, 2014, 55(1): 101-115.

[12] LIU D, LI T R, HU P, et al. Multiple-Category Classification with Decision-Theoretic Rough Sets[M] // Rough Set and Knowledge Technology. Springer Berlin Heidelberg, 2010: 703-710.

[13] ZHANG Z, WANG R. Applying Three-way Decisions to Sentiment Classification with Sentiment Uncertainty[M] // Rough Sets and Knowledge Technology. Springer International Publishing, 2014: 720-731.

[14] JIA X Y, LIAO W, TANG Z, et al. Minimum cost attribute reduction in decision-theoretic rough set models[J]. Information Sciences, 2013, 219: 151-167.

[15] LI H X, ZHOU X, ZHAO J, et al. Non-Monotonic Attribute Reduction in Decision-Theoretic Rough Sets[J]. Fundamenta Informaticae, 2013, 126(4): 415-432.

[16] JIA X Y, ZHENG K, LI W W, et al. Three-Way Decisions Solution to Filter Spam Email: An Empirical Study[M]// Rough Sets and Current Trends in Computing. Springer Berlin Heidelberg, 2012: 287-296.

[17] ZHOU B, YAO Y Y, LUO J G. A Three-Way Decision Approach to Email Spam Filtering[C]// Advances in Artificial Intelligence, 23rd Canadian Conference on Artificial Intelligence (AI 2010). Ottawa, Canada, 2010: 28-39.

[18] LI F, MIAO D Q, LIU C H, et al. Image segmentation decision rough set[J]. CAAI Transactions on Intelligent System, 2014 (2): 143-147. (in Chinese)
李峰，苗夺谦，刘财辉，等. 基于决策粗糙集的图像分割[J]. 智能系统学报，2014(2): 143-147.

[19] LI X Y, ZHANG Q Q. An image classification algorithm based on compatible granular model and three decision making[J]. Computing Technology and Automation, 2014, 33(4): 93-96. (in Chinese)
李晓艳，张倩倩. 基于相容粒模型和三支决策的图像分类算法[J]. 计算技术与自动化，2014, 33(4): 93-96.

[20] JIA X Y, SHANG L. A Simulated Annealing Algorithm for Learning Thresholds in Three-way Decision-theoretic Rough Set Model[J]. Journal of Chinese Computer Systems, 2013, 34(11): 2603-2606. (in Chinese)
贾修一，商琳. 一种求三支决策阈值的模拟退火算法[J]. 小型微型计算机系统，2013, 34(11): 2603-2606.

[21] LIN F T, KAO C Y, HSU C C. Applying the genetic approach to simulated annealing in solving some NP-hard problems[J]. IEEE Transactions on Systems Man & Cybernetics, 1993, 23(6): 1752-1767.

(下转第 61 页)

表 3　各种算法的平均 RI 值

数据集	CSPA	HGPA	HGBF	EMcN	MCLA	NBKCE
2D4C	0.9478	0.6271	0.9581	**0.9708**	**0.9708**	**0.9708**
wine	0.6836	0.5379	**0.7247**	0.7187	0.7187	0.7194
iris	**0.8859**	0.5416	0.8846	0.8797	0.8797	0.8797
glass	0.7233	0.6481	0.7155	0.6975	0.7054	**0.7234**
segment	0.8328	0.7554	0.8362	0.8369	0.8296	**0.8373**
balance	0.5780	0.5339	0.5825	0.5845	0.5919	**0.6271**
diabetes	0.4886	0.4994	0.5082	**0.5507**	**0.5507**	**0.5507**
heart	0.5079	0.5014	0.5097	0.5041	0.5041	**0.5141**
liver	0.5071	0.4989	0.5012	0.5012	0.5012	**0.5043**
cmc	0.5575	0.5484	**0.6374**	0.5577	0.5580	0.5582
最优次数	1	0	2	2	2	7
最劣次数	1	9	0	1	1	0

在 F1 评价指标下，NBKCE 算法在 10 个数据集中取得了 7 次最优结果，0 次最劣结果，相较于对比算法有效地提高了聚类集成的质量；在 R2 评价指标下，其取得了 7 次最优结果，在其他 3 个数据集中也取得了较好的结果。由此可以看出，NBKCE 能够有效使用数据集的潜在信息来提高聚类质量。

结束语　本文提出了一种基于非负矩阵分解(NMF)的聚类集成方法 NBKCE。该算法将来自于原数据集的关系矩阵与信息矩阵结合后融入到共识函数中，利用 NMF 技术获取隶属度矩阵，有效利用潜在信息，提高了聚类集成的性能。今后的工作将考虑加入部分监督信息以改善集成效果，同时考虑将该聚类集成算法应用于多视图聚类集成的研究，以提高多视图聚类的性能。

参 考 文 献

[1] YANG C Y, LIU D Y, YANG B, et al. The research on clustering ensemble[J]. Computer Science, 2011, 38(2): 166-170. (in Chinees)
杨草原，刘大有，杨博，等. 聚类集成方法研究[J]. 计算机科学，2011，38(2)：166-170.

[2] STREHL A, GHOSH J. Cluster ensembles-a knowledge reuse framework for combining multiple partitions[J]. Journal of Machine Learning Research, 2003, 3(3): 583-617.

[3] WANG H J, LI Z S, CHENG Y, et al. A Latent Variable Model for Cluster Ensemble[J]. Journal of Software, 2009, 20(4): 825-833. (in Chinees)
王红军，李志蜀，成飚，等. 基于隐含变量的聚类集成模型[J]. 软件学报，2009，20(4)：825-833.

[4] ZHOU Z H. Ensemble Methods: Foundations and Algorithms [M]. Taylor & Francis, 2012.

[5] YANG Y, KAMEL M. An aggregated clustering approach using multi-ant colonies algorithms[J]. Pattern Recognition, 2006, 38(7): 1278-1289.

[6] LAMON N, BOONGOEN T, GARRETT S. Link-based cluster ensemble approach for categorical data clustering[J]. IEEE Transactions on Knowledge and Data Engineering, 2012, 24(3): 413-425.

[7] YANG Y, FENG C F, JIA Z, et al. A Link-Based Fuzzy Clustering Ensemble[J]. Journal of University of Electronic Science and Technology of China, 2014, 43(6): 887-892. (in Chinees)
杨燕，冯晨菲，贾真，等. 基于链接的模糊聚类集成方法[J]. 电子科技大学学报，2014，43(6)：887-892.

[8] HAN J, KAMBER M. Data Mining: Concepts and Techniques [J]. Data Mining Concepts Models Methods & Algorithms Second Edition, 2006, 5(4): 1-18.

[9] DING C, HE X, SIMON H. Nonnegative lagrangian relaxation of k-means and spectral clustering[C]// ECML. 2005: 530-538.

[10] ZHANG J S, WANG C P, YANG Y Q. Learning latent features by nonnegative matrix factorization combining similarity judgments[J]. Neurocomputing, 2015, 155: 43-52.

[11] MIAO L D, QI H R. Endmember extraction from highly mixed data using minimum volume constrained nonnegative matrix factorization[J]. IEEE Trans. Geosci. Remote Sens., 2007, 45(3): 765-777.

[12] ASUNCION A, NEWMAN D J. UCI machine learning repository school of information and computer science, university of california [DB/OL]. (2007-06-02). http://www.ics.uci.edu/-mlearn/MLRepository.html.

[13] FERN X Z, BRODLEY C E. Solving cluster ensemble problems by bipartite graph partitioning[C]// Proc. 21th Int. Conf. Mach. Learn.. 2004: 36-44.

[14] ALEXANDER T, ANIL K J, WILLIAM P. Clustering Ensembles: Models of Consensus and Weak Partitions[J]. IEEETrans. on Pattern Analysis and Machine Intelligence, 2005, 27(12): 1866-1881.

[15] YANG Y, JIN F, KAMEL M. Survey of clustering validity evaluation[J]. Application Research of Computer, 2008, 25(6): 1630-1632. (in Chinese).
杨燕，靳蕃，KAMEL M. 聚类有效性评价综述[J]. 计算机应用研究，2008，25(6)：1630-1632.

[16] RAND W M. Objective criteria for the evaluation of clustering methods[J]. Journal of American Statistical Association, 1971, 66(336): 846-850.

(上接第 44 页)

[22] YAO Y Y. Interval sets and interval-set algebras[C]// IEEE International Conference on Cognitive Informatics (ICCI 2009). Hong Kong, China, 2009: 307-314.

[23] ZADEH L A. Fuzzy sets* [J]. Information & Control, 1965, 8(3): 338-353.

[24] PEDRYCZ W. Shadowed sets: representing and processing fuzzy sets[J]. IEEE Transactions on Systems Man & Cybernetics Part B Cybernetics A Publication of the IEEE Systems Man & Cybernetics Society, 1998, 28(1): 103-109.

[25] ABDULLAH S, GOLAFSHAN L, ZAKREE M, et al. Re-heat simulated annealing algorithm for rough set attribute reduction [J]. International Journal of Physical Sciences, 2011(8): 2083-2089.

[26] CHEN Y P, EAMONN K, HU B, et al. The UCR Time Series Classification Archive [DB/OL]. http://www.cs.ucr.edu/~eamonn/time_series_data.