基于 *K*-Shape 的时间序列模糊分类方法

李海林[1,2*]，贾瑞颖[1]，谭观音[1,2]

(1. 华侨大学信息管理与信息系统系 福建 泉州 362021；2. 华侨大学应用统计与大数据研究中心 福建 厦门 361021)

【摘要】时间序列分类是数据挖掘中的重要主题，现有的大部分时间序列分类方法较少考虑到序列形状对分类结果的影响。该文提出了一种基于 *k*-shape 的时间序列模糊分类方法。该方法通过使用 *k*-shape 聚类算法对时间序列训练数据集各类别的成员进行聚类，获得各类别的聚类中心并形成聚类中心群，将每个类别的聚类中心群作为时间序列数据模糊分类的初始聚类中心，根据隶属度最大原则确定测试时间序列数据的类别标签。在 30 个时间序列公开数据集上的分类实验结果表明，该方法相较于 SVM、Bayes、EAIW 和 TLCS 这 4 种分类算法具有更好的分类性能，对具有扭曲和位移特征的时间序列数据分类有更好的可用性。

关 键 词 分类算法； 模糊分类； *k*-shape； 时间序列

中图分类号 TP273 **文献标志码** A doi:10.12178/1001-0548.2020380

Fuzzy Classification for Time Series Data Based on *K*-Shape

LI Hailin[1,2*], JIA Ruiying[1], and TAN Guanyin[1,2]

(1. Department of Information Management and Information Systems, Huaqiao University Quanzhou Fujian 362021;
2. Research Center for Applied Statistics and Big Data, Huaqiao University Xiamen Fujian 361021)

Abstract Time series classification is an important topic in data mining. Most existing time series classification methods do not consider the influence of the shape of the time series on the classification results. The paper proposes a fuzzy classification method for time series based on *k*-shape. The method utilizes the *k*-shape clustering algorithm to cluster each category of the time series training datasets and obtains the cluster centers group of each class. After utilizing the cluster center group of each class as the initial clustering center of the fuzzy classification, class labels of the test datasets are determined according to the principle of maximum membership degree. Experimental results on 30 time series public datasets show that the proposed method has better classification performance than the traditional methods, including support vector machine (SVM), Bayes, ensemble algorithm of interval weightsc (EAIW), and trend information based on longest common subsequence (TLCS), with more excellent usability for time series with distortion and displacement characteristics.

Key words classification algorithm; fuzzy classification; *k*-shape; time series

时间序列是一种与时间相关的数值型数据，基于时间序列的数据挖掘与分析成为目前数据研究领域中最具有挑战性的十大问题之一[1]。分类算法是时间序列数据挖掘中极为重要的任务和技术[2]，有大量关于时间序列分类和挖掘的研究[3]。分类问题依赖于时间序列间的相似性度量，而相似性度量是两条时间序列相似程度的度量方法[4]。对于时间序列来说，同类时间序列间的相似性主要有时域相似性、形状相似性和变化相似性 3 种形式[5]。支持向量机 (support vector machine, SVM) 是由文献 [6] 提出的通过核函数将时间序列向高维空间映射的方法，可用于时间序列分类。朴素贝叶斯分类器是目前公认的一种简单而有效的概率分类方法，作为经典的机器学习算法之一，在信息检索领域有着极为重要的地位[7]。文献 [8] 提出了 EAIW 分类算法，该算法为时间序列区间赋予权值，采用集成分类的算法，通过权值对时间序列进行分类。文献 [9] 提出了 TLCS 算法，该算法提出一种新的基于时间序列的趋势离散化方法，利用 LCS 对其进行相似性度量。

收稿日期：2020－10－10；修回日期：2021－06－25
基金项目：国家自然科学基金面上项目 (71771094)；福建省自然科学基金面上项目 (2019J01067)；福建省社会科学规划一般项目 (FJ2020B088)
作者简介：李海林 (1982－)，男，教授，博士生导师，主要从事数据挖掘与决策支持方面的研究.
*通信作者：李海林，E-mail：blihailin@163.com

模糊聚类由于能够描述样本类属的中介性，能更客观地反映现实世界，目前已成为聚类分析的主流，成为非监督模式识别的一个重要分支。模糊聚类分析已经成功地应用于遥感图像处理、医学图像处理、基因数据处理、模糊决策分析等领域[10]。众多的模糊聚类方法中，应用最广泛的是模糊 C-均值 (fuzzy C-means, FCM) 算法。由于模糊 C-均值算法在初始聚类中心的选择上具有随机性，对初始值比较敏感，难以取得全局最优[11]。

本文提出一种新的时间序列分类方法，即基于 k-shape 的时间序列模糊分类方法。该方法利用一种新的k-shape 聚类算法[12] 对训练集中每个类的时间序列进行聚类，得到聚类中心群，并将这些中心群作为模糊分类的初始聚类中心，使用模糊 C 均值对时间序列测试集数据进行分类。

1 相关理论基础

k-shape 是一种应用在时间序列数据中的聚类方法[12]，此算法提出基于时间序列形态相似性的距离量度方式 SBD，并采用一种新的聚类中心计算方式 SE 提取每类聚类中心的时间序列曲线形态，以此完成聚类。

定义 1 SBD 是一种基于形状的相似性度量方式，在一定程度上弥补了以欧式距离作为相似性评价指标的不足，通过 SBD 可得到两条时间序列之间的相似度量。具体过程如下：

基于形状的相似性度量方法：$(\mathrm{dist}, y') = \mathrm{SBD}(x, y)$

输入：两条 Z-score 标准化后的时间序列 x，y；

输出：时间序列 x，y 的相似性度量 dist 和y相对于x的对齐序列y'；

计算$\mathrm{len} = 2^{2*\mathrm{length}(x)-1}$；

对x和y分别进行快速傅里叶变换，即 $F_x = \mathrm{FFT}(x, \mathrm{len})$，$F_y = \mathrm{FFT}(y, \mathrm{len})$；

进行逆快速傅里叶变换，计算x和y之间的交相关序列 CC，$\mathrm{CC} = \mathrm{IFFT}\{F_x * F_y\}$；

根据系数归一化思想，计算 $[\mathrm{value}, \mathrm{index}] = \frac{\mathrm{CC}}{\|x\|\|y\|}$，value 为最大值，index 为此时序列$x$和$y$对齐的起始位置；

$\mathrm{dist} = 1 - \mathrm{value}$，$\mathrm{shift} = \mathrm{index} - \mathrm{length}(x)$；

判断 shift 大小。若 $\mathrm{shift} \geqslant 0$，则 $y' = [\mathrm{zeroes}(1, \mathrm{shift}), y(1 : \mathrm{end} - \mathrm{shift})]$，否则 $y' = [y\ (1-\mathrm{shift}: \mathrm{end}), \mathrm{zeroes}(1, -\mathrm{shift})]$。

由 SBD 算法可得到时间序列x和y之间的相似度量 dist。当 dist 为 0 时，表示两条序列完全相似。同时，根据对齐的起始位置 index，可得到y相对于x的对齐序列y'。

定义 2 SE 是从时间序列中提取最具有代表的形态特征，以此进行聚类。根据文献 [12]，k-shape 利用 SE 方法可以在每种类别的数据中产生一个聚类中心。

基于代表性形态特征的聚类中心提取方法：$\boldsymbol{C} = \mathrm{SE}(\boldsymbol{X}, \boldsymbol{R})$

输入：Z-score 标准化后的时间序列集合 $\boldsymbol{X} = [x_1, x_2, \cdots, x_n]$ (其中每条时间序列的长度为p)，与 $\boldsymbol{X}$ 对齐的参考序列向量$\boldsymbol{R}$；

输出：聚类中心矩阵$\boldsymbol{C}$；

设立对齐序列矩阵$\boldsymbol{M}$；

对 $\boldsymbol{X}$ 中的每条时间序列x_i，计算$[\mathrm{dist}, x_i'] = \mathrm{SBD}(\boldsymbol{R}, x_i)$，并将$x_i'$加入$\boldsymbol{M}$；

提取$\boldsymbol{M}$的主要特征向量，$\boldsymbol{S} = \boldsymbol{M}^{\mathrm{T}}\boldsymbol{M}$，$\boldsymbol{Q} = \boldsymbol{I} - \frac{1}{p}\boldsymbol{O}$ ($\boldsymbol{I}$为单位矩阵，$\boldsymbol{O}$为全 1 矩阵)，$\boldsymbol{M} = \boldsymbol{Q}^{\mathrm{T}}\boldsymbol{S}\boldsymbol{Q}$；

提取第 1 特征向量，即$\boldsymbol{C} = \mathrm{Eig}(\boldsymbol{M}, 1)$。

定义 3 k-shape 是一种基于时序形状的时间序列聚类方法。该方法首先利用 SBD 算法进行时间序列之间的相似性度量，获得相似序列。然后使用 SE 算法从相似序列中提取第一特征向量，作为聚类中心，进而完成聚类。

基于时间序列形态特征的聚类方法：$(\mathbf{IDX}, \boldsymbol{C}) = k\text{−shape}(\boldsymbol{X}, k)$

输入：聚类中心数k和 Z-score 标准化后的同类别的时间序列集$\boldsymbol{X} = [x_1, x_2, \cdots, x_n]$，其中$x_i$表示某一条时间序列；

输出：时间序列集$\boldsymbol{X}$的类别标签向量 **IDX** 和时间序列集$\boldsymbol{X}$的聚类中心矩阵$\boldsymbol{C}$；

设置初始迭代次数 iter=0，类别标签向量 **IDX** 为空向量；

while iter < 100 do

for $j \leqslant k$ do

创建一个新的时间序列矩阵$\boldsymbol{X}'$

for $i \leqslant n$ do

if $\mathbf{IDX}(i) = j$ then

$\boldsymbol{X}' = [\boldsymbol{X}', \boldsymbol{X}(i)]$

$\boldsymbol{C}(j) = \mathrm{SE}(\boldsymbol{X}', \boldsymbol{C}(j))$

for $i \leqslant n$ do

设置$\min \mathrm{dist} = \infty$

for $j \leqslant k$ do

$[\mathrm{dist}, x'] = \mathrm{SBD}(\mathbf{C}(j), \mathbf{X}(i))$

if dist < mindist then

IDX(i) = j

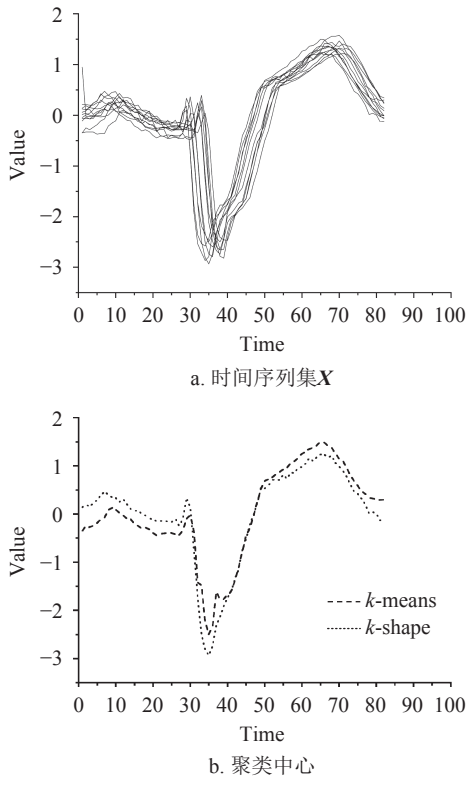
iter = iter + 1

2　时间序列模糊分类

新分类方法首先通过基于 *k*-shape 的聚类中心群方法构建每个类别的聚类中心群，然后结合基于 FCM 的模糊分类方法实现对时间序列的分类。该方法具有良好的分类性能。

2.1　基于 *k*-shape 的聚类中心群

绝大多数的时间序列都存在明显的位移和扭曲，传统的聚类算法不能有效解决这部分时间序列的聚类问题，而 *k*-shape 对具有位移和扭曲的时间序列聚类有更好的适用性，可以在一定程度上弥补传统聚类算法以欧氏距离作为相似性度量指标的不足。本文提出一种新的基于 *k*-shape 的聚类中心群方法 KCG，该方法可得到单个类别的聚类中心群，且有较好的代表性。



a. 时间序列集$\boldsymbol{X}$

b. 聚类中心

图 1　*k*-shape 算法优势

如图 1 所示，通过对时间序列数据集 $\boldsymbol{X}$ 提取聚类中心，可以看出 *k*-shape 提取的聚类中心相比于传统聚类算法 *k*-means 更符合数据集 $\boldsymbol{X}$ 的形态特征，更具有代表性。其算法过程如下。

基于 *k*-shape 聚类的聚类中心群方法：$\boldsymbol{C}_a$ = KCG($\boldsymbol{X}$,k)

输入：同一类别的时间序列数据集$\boldsymbol{X}=[x_1, x_2,\cdots,x_n]$、$\boldsymbol{X}$ 的类别标签 A 和聚类中心数k；

输出：聚类中心矩阵$\boldsymbol{C}_a=[c_{1a},c_{2a},\cdots,c_{ka}]$，$c_{ia}$ 表示A类别的聚类中心群中的第 i 个中心代表对象；

1) 根据 *k*-shape 聚类算法，将时间序列数据集 X 划分成k类，得到 k 个聚类中心，即($\mathbf{IDX}_a$,C_a) = *k*-shape($\boldsymbol{X}$,k)，$\mathbf{IDX}_a$和$\boldsymbol{C}_a$分别代表 $\boldsymbol{X}$ 中所有序列的聚类标签向量和聚类中心矩阵；

2) 将步骤 1) 中得到的聚类中心矩阵$\boldsymbol{C}_a$标记为$\boldsymbol{C}_a=[c_{1a},c_{2a},\cdots,c_{ka}]$，代表$A$类别成员序列的聚类中心群。

通过基于 *k*-shape 聚类的聚类中心群方法 KCG 可以得到同类别时间序列集的聚类中心群，该中心群可代表整个类别的时间序列数据形态特征分布情况。

2.2　模糊分类

模糊分类相比于传统分类算法的硬划分，更符合时间序列分类的不确定性。FCM 算法作为模糊分类的主流算法之一，能在一定程度上解决时间序列数据分类的不确定性问题，是传统硬聚类算法的一种改进。

FCM 算法的核心思想是通过极小化目标函数求最优聚类中心[13]，聚类结果是每一条时间序列对聚类中心的隶属程度，该隶属程度用一个数值来表示。本文提出一种基于 FCM 的模糊分类方法，通过已知的初始聚类中心群，进行模糊分类，降低初始值对最后分类结果的影响。为了便于理解和讨论模糊分类方法，假设时间序列数据集共分为两类，进一步解释该算法。其具体算法过程如下：

基于 FCM 的模糊分类算法：$\boldsymbol{D}$=FCM ($\boldsymbol{Y}$,$\boldsymbol{C}$, iter_max)

输入：包含已知类别A和类别B的聚类中心矩阵$\boldsymbol{C}$、允许的最大迭代次数iter_max(默认为 100) 和时间序列测试数据集$\boldsymbol{Y}=[y_1,y_2,\cdots,y_m]$；

输出：隶属度矩阵$\boldsymbol{D}$；

1) 将聚类中心矩阵$\boldsymbol{C}$中A类别的聚类中心群标记为$\boldsymbol{C}_a$，B类别的聚类中心群标记为$\boldsymbol{C}_b$；

2) 根据 FCM 模糊聚类算法，将$\boldsymbol{C}$作为初始聚类中心进行聚类，得到模糊隶属度矩阵$\boldsymbol{D}$，即$\boldsymbol{D}$ = FCM($\boldsymbol{Y}$,$\boldsymbol{C}$,iter_max)。

通过基于模糊分类的 FCM 算法得到模糊隶属度矩阵$\boldsymbol{D}$后，进一步分别计算y_i属于$\boldsymbol{D}$中A类别聚类中心群$\boldsymbol{C}_a$和B类别聚类中心群$\boldsymbol{C}_b$的隶属度之和，并判断大小。较大的隶属度之和代表的类标签为y_i所属类别，即可完成分类。

基于 FCM 的模糊分类算法以 *k*-shape 聚类得到的聚类中心群作为初始聚类中心，经过一定次数的迭代，得到模糊隶属度矩阵，依次判断测试时间序列属于各类别标签的聚类中心群的隶属度之和，根据最大隶属度原则确定该测试时间序列属于哪个类别，从而完成时序数据的分类。

2.3 基于 *k*-shape 的时间序列模糊分类方法

考虑到 *k*-shape 聚类算法和模糊分类算法的优势，本文将基于 *k*-shape 的聚类中心群算法 KCG 和基于 FCM 的模糊分类算法结合起来，提出了一种思路更为简单的时间序列分类方法 (*k*-shape and FCM based time series clustering, KFCM)。KFCM 算法首先将时间序列训练集各类别的序列成员进行 *k*-shape 聚类，分别得到每个类别的聚类中心群，形成已知类标签的聚类中心群。然后，使用基于 FCM 的模糊分类算法，将测试集序列与已知标签的聚类中心群进行聚类，输出模糊隶属度矩阵。最后，根据隶属度大小原则进一步判断测试集类别。具体算法如下：

基于 *k*-shape 的时间序列模糊分类方法：$\boldsymbol{L}=\mathrm{KFCM}(\boldsymbol{X},\boldsymbol{Y},k,\mathrm{iter\_max})$。

输入：训练集 $\boldsymbol{X}$、测试集 $\boldsymbol{Y}$、默认最大迭代次数 iter_max 和聚类中心数 k；

输出：测试集 $\boldsymbol{Y}$ 的类标签向量 $\boldsymbol{L}$；

1) 根据训练集 $\boldsymbol{X}=[x_1,x_2,\cdots,x_n]$ 的成员类标签 $[h_1,h_2,\cdots,h_w]$，依次利用 KCG 对类别 h_i 包含的每个成员进行聚类，得到 h_i 的聚类中心群 $\boldsymbol{C}_i$。将 w 个聚类中心群合并得到总聚类中心群 $\boldsymbol{C}$，$\boldsymbol{C}$ 包含 w 个类别，共 kw 个聚类中心；

2) 对测试集 $\boldsymbol{Y}=[y_1,y_2,\cdots,y_m]$，使用基于 FCM 的模糊分类算法，即 $\boldsymbol{D}=\mathrm{FCM}(\boldsymbol{Y},\mathbf{C},\mathrm{iter\_max})$；

3) 根据步骤 2) 得到的模糊隶属度矩阵 $\boldsymbol{D}$，计算测试集对象 y_j 属于 $\boldsymbol{D}$ 中各类聚类中心群的隶属度之和，其较大者的类标签为 y_j 所属类别 l_j；

4) 重复步骤 3)，获得 $\boldsymbol{Y}$ 中所有成员的类标签，即 $\boldsymbol{L}=[l_1,l_2,\cdots,l_m]$，其中 m 为测试集 $\boldsymbol{Y}$ 包含的时间序列数目。

3 数值实验与结果分析

为了验证本文提出算法的有效性，在 30 个时间序列数据集上做分类实验。通过实验结果可以验证分类精度的有效性，也可以验证针对存在位移和扭曲特征的时间序列分类，新方法的适用性。

3.1 实验设置

算法代码使用 Python 3.7 在 Anaconda 科学计算环境中实现，运行所用计算机的 CPU 型号为 InterCore i5-8250U (1.60 GHz)，RAM 为 16 GB，操作系统是 Windows10 64 位 (DirectX 12)。

本文采用的数据集是 UCR TS Archive 2015，UCR[14] 是时间序列数据集，每个数据集样本都带有样本类别标签，它是目前时间序列挖掘领域重要的开源数据集资源。从 UCR 数据集中选取了 30 个训练集，为了验证新方法具有更高的分类质量和性能，这 30 个数据集在类别、长度和大小上具有明显差异。具体数据集信息如表 1 所示。

表 1 时间序列数据集

DataSet	No. of Class	TrainSize	TestSize	Length
Cricket_Z	12	390	390	300
Cricket_X	12	390	390	300
ToeSegmentation1	2	40	228	277
ToeSegmentation2	2	36	130	343
DistalPhalanxTW	6	139	400	80
Ham	2	109	105	431
ProximalPhalanx-TW	6	205	400	80
ECGFiveDays	2	23	861	136
DistalPhalanxOutlineAgeGroup	3	139	400	80
Haptics	5	155	308	1 092
PhalangesOutlinesCorrect	2	1 800	858	80
DiatomSizeReduction	4	16	306	345
ProximalPhalanxOutlineCorrect	2	600	291	80
TwoLeadECG	2	23	1 139	82
SonyAIBORobotSurface1	2	20	601	70
FacesFOUR	4	24	88	350
InlineSkate	7	100	550	1 882
MiddlePhalanxTW	6	154	399	80
BeetleFly	2	20	20	512
ShapeletSim	2	20	180	500
WormsClass	5	77	181	900
MoteStrain	2	20	1 252	84
Oliveioil	4	30	30	570
trace	4	100	100	275
beef	5	30	30	470
ECG200	2	100	100	96
Symbols	6	25	995	398
Coffee	2	28	28	286
Car	4	60	60	577
MiddlePhalanxOutlineCorrect	2	291	600	80

在对训练数据集进行 *k*-shape 聚类时，训练集的类别和类别数是已知的，需要用 *k*-shape 单独对训练集每一个类别包含的时间序列进行聚类，进而确定各类别的聚类中心群。

3.2 参数设置

在进行参数讨论时，假定训练集共有 w 个类别，每个类别的聚类中心数都是 k，因此共可得到

wk 个聚类中心，且每个聚类中心群都标记着本类别的标签。本文以 Cricket_X 数据集为例，说明利用手肘法选取最佳聚类数 *k* 的过程。*k* 的取值为 1～8 (本文设置上限为 8)。如图 2 所示，随着 *k* 增大，SSE 的下降幅度会骤减，在 *k*=4 之后下降幅度趋于平缓。因此针对 Cricket_X 数据集，最佳聚类中心数是 4。

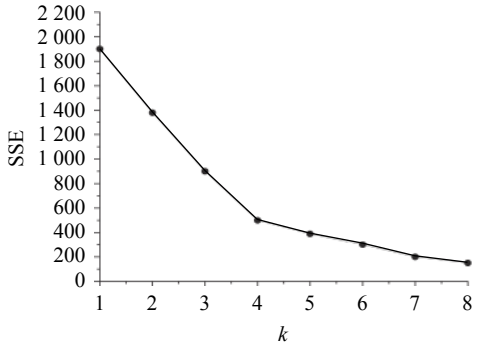


图 2　*k*值选取对 SSE 的影响

3.3　数据分类

本节将提出的 KFCM 算法与 SVM[6]、Bayes[7]、EAIW[8] 和 TLCS[9] 这 4 种基准分类算法进行比较。为了进一步验证 *k*-shape 在提取聚类中心过程中的算法优势，本文利用 *k*-means 算法提取各类别的聚类中心群来作为模糊分类的初始聚类中心，同时使用模糊分类方法对时间序列数据集进行分类。该算法称为 KMFCM，作为 KFCM 的对比算法之一。利用以上 6 种方法对 30 组 UCR 数据集进行分类实验，分类错误率如表 2 所示。利用平均分类错误率、方差和胜出率 3 个指标评价分类效果，如表 3 所示。

从表 2 可知，KFCM 算法的平均错误率最低，错误率的总体波动幅度较小，在 30 个数据集中有 9 个数据集的错误率最小，有最高的胜出率。为了更好地表现分类结果，本文通过对错误率的两两比较，使用可视化分类比较结果展示，如图 3 所示。KFCM 的分类准确率和胜出率都高于 KMFCM，同时，在 ShapeletSim、MoteStrain、InlineSkate 等数据集上也有较低的分类错误率，这证明了 *k*-shape 相比于 *k*-means 在聚类过程中有明显优势。KFCM 在时间序列测试集上的分类性能优于 SVM 和 Bayes，平均准确率有一定提升。最后，KFCM 分类准确率相较于 TLCS 和 EAIW 也有一定提高，个别数据集提高效果较为明显。

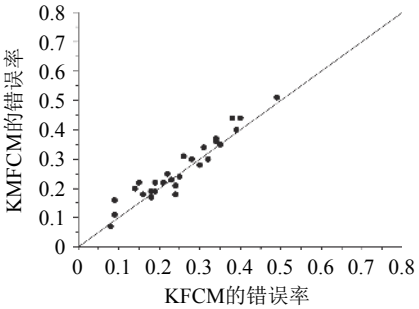
表 2　各分类算法分类错误率

DataSet	分类错误率					
	KFCM	KMFCM	SVM	Bayes	EAIW	TLCS
Cricket_Z	0.31	0.34	0.50	0.60	**0.22**	0.52
Cricket_X	0.39	0.40	0.48	0.61	**0.22**	0.58
ToeSegmentation1	**0.15**	0.22	0.38	0.44	0.16	0.23
ToeSegmentation2	0.16	0.18	0.26	0.36	**0.09**	0.18
DistalPhalanxTW	0.24	**0.18**	0.22	0.24	0.35	0.46
Ham	0.28	0.30	**0.25**	0.26	0.43	0.38
ProximalPhalanxTW	**0.19**	**0.19**	0.20	0.21	0.20	0.27
ECGFiveDays	0.18	0.17	0.16	0.22	**0.11**	0.15
DistalPhalanxOutlineAgeGroup	0.21	0.22	0.18	**0.17**	0.26	0.38
Haptics	**0.49**	0.51	0.54	0.56	0.61	0.54
PhalangesOutlinesCorrect	0.34	0.36	**0.21**	0.34	0.25	0.31
DiatomSizeReduction	**0.09**	0.11	0.10	0.21	0.31	0.14
ProximalPhalanxOutlineCorrect	0.24	0.21	**0.14**	0.35	0.19	0.24
TwoLeadECG	0.25	0.24	0.28	0.30	**0.19**	0.26
SonyAIBORobotSurface1	0.09	0.11	0.22	**0.06**	0.30	0.23
FacesFour	0.22	0.25	0.36	**0.10**	0.20	0.19
InlineSkate	**0.40**	0.44	0.75	0.51	0.51	0.53
MiddlePhalanxTW	**0.35**	**0.35**	0.37	0.40	0.45	0.51
BeetleFly	0.21	0.22	**0.15**	0.20	0.25	**0.15**
ShapeletSim	**0.14**	0.20	0.53	0.15	0.31	0.27
WormsClass	**0.34**	0.37	0.59	0.42	0.40	0.38
MoteStrain	**0.09**	0.16	0.15	0.17	0.13	**0.09**
Oliveioil	0.19	0.22	0.17	**0.07**	0.39	0.50
Trace	0.30	0.28	0.29	**0.20**	0.21	0.26
Beef	0.26	0.31	0.27	0.37	0.19	**0.10**
ECG200	0.18	0.19	**0.16**	0.33	0.27	0.19
Symbols	0.32	0.30	0.24	0.36	0.08	**0.02**
Coffee	0.08	0.07	0.05	0.07	0.18	**0.04**
Car	0.23	0.23	0.30	0.42	0.27	**0.17**
MiddlePhalanxOutlineCorrect	0.38	0.44	0.45	0.46	**0.26**	0.37

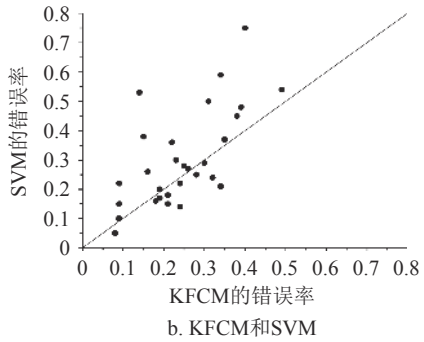
表 3　各类算法评价指标

参数	KFCM	KMFCM	SVM	Bayes	EAIW	TLCS
mean	**0.25**	0.26	0.30	0.31	0.27	0.29
variance	**0.01**	**0.01**	0.03	0.02	0.02	0.03
win rate	**0.30**	0.10	0.17	0.17	0.20	0.20

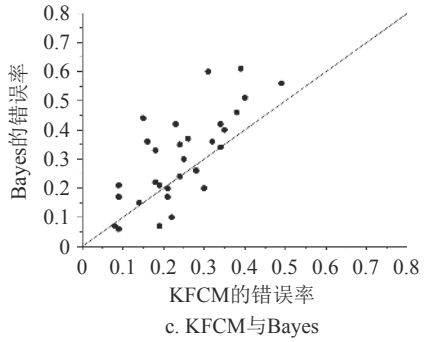
本文选取 4 个时间序列数据集，来说明聚类中心数*k*对分类结果的影响，如图 4 所示。*k* 值对各数据集的分类结果影响不同，对于 BeetleFly 数据集，错误率最大值和最小值之间差距达到 0.23；对于 OliveOil 数据集，错误率最大值和最小值之间差距为 0.06。



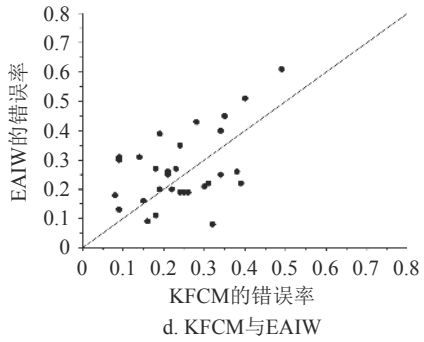
a. KFCM和KMFCM



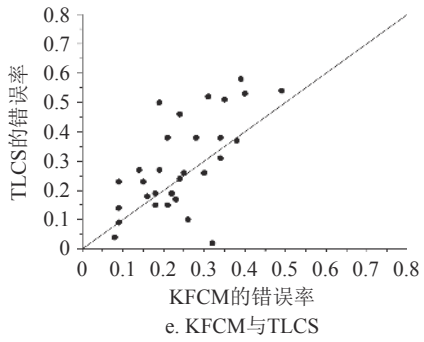
b. KFCM和SVM



c. KFCM与Bayes



d. KFCM与EAIW



e. KFCM与TLCS

图 3 分类算法错误率的比较

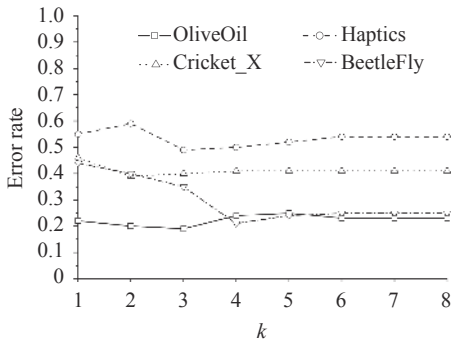


图 4 k与错误率的关系

由此可知，对于不同的数据集，k 对最后分类效果的影响也是不同的。有的数据集如 BeetleFly 受影响比较大，因此参数 k 需根据具体数据而设定。

3.4 实验分析

本文分析了 30 个时间序列训练集中部分时间序列数据的特点，将时间序列训练集按照是否存在明显位移和扭曲的特点分为两大类，计算每一类的平均错误率。横向比较，新方法 KFCM 在存在明显位移和扭曲特点的时间序列平均错误率要比趋势较为一致的时间序列平均错误率要低；纵向比较，对于存在明显扭曲和位移的时间序列集，新方法 KFCM 相比其他 5 个方法的错误率低，分类性能更好。

从图 5 中看出，InlineSkate、MoteStrain 和 ShapeletSim 这 3 个时间序列数据集具有较明显的位移和扭曲，故 SVM、Bayes 和 KMFCM 在处理此类的数据集分类时存在较大的不足，但 KFCM 算法表现良好。Symbols、TwoLeadECG 和 Car 这 3 个时间序列集中同一类时间序列数据上“几乎”不存在位移或者扭曲，故 SVM、Bayes 和 KMFCM 在这 3 个数据集上的分类效果比较理想。图 6 也表明新方法 KFCM 在 InlineSkate、MoteStrain 和 ShapeletSim 时间序列数据集上有更低的错误率。

3.5 时间效率比较

时间序列分类的性能不仅体现在分类效果上，而且也包含了算法的整体时间消耗。为了更直观的比较以上 6 种分类方法的时间效率，本文随机抽取了 WormsTwoClass、Ham、Beef 和 FaceFour 这 4 个时间序列数据集，进行方法运行时间效率的对比实验。

由图 7 可知，KFCM 方法在 4 个时间序列数据集上的时间效率比 KMFCM、SVM 和 Bayes 低，分类时间较长。但相比于 EAIW 和 TLCS，KFCM 方法所花费的时间则相对较少。总体来讲，KFCM 的时间效率要高于 EAIW 和 TLCS，低于 SVM、Bayes 和 KMFCM。文献 [15] 在大型电力负荷时间序列曲线聚类实验中证明 k-shape 的时间复杂度高于 k-means。本文研究的时间序列分类方法主要是通过 k-shape 提取聚类中心群，该阶段的计算复杂度较高，导致整体算法在大数据环境下的时间效率较低，并不适用于大型数据集。但结合分类质量来看，KFCM 可以使用较高的时间效率来获得较好的分类效果。

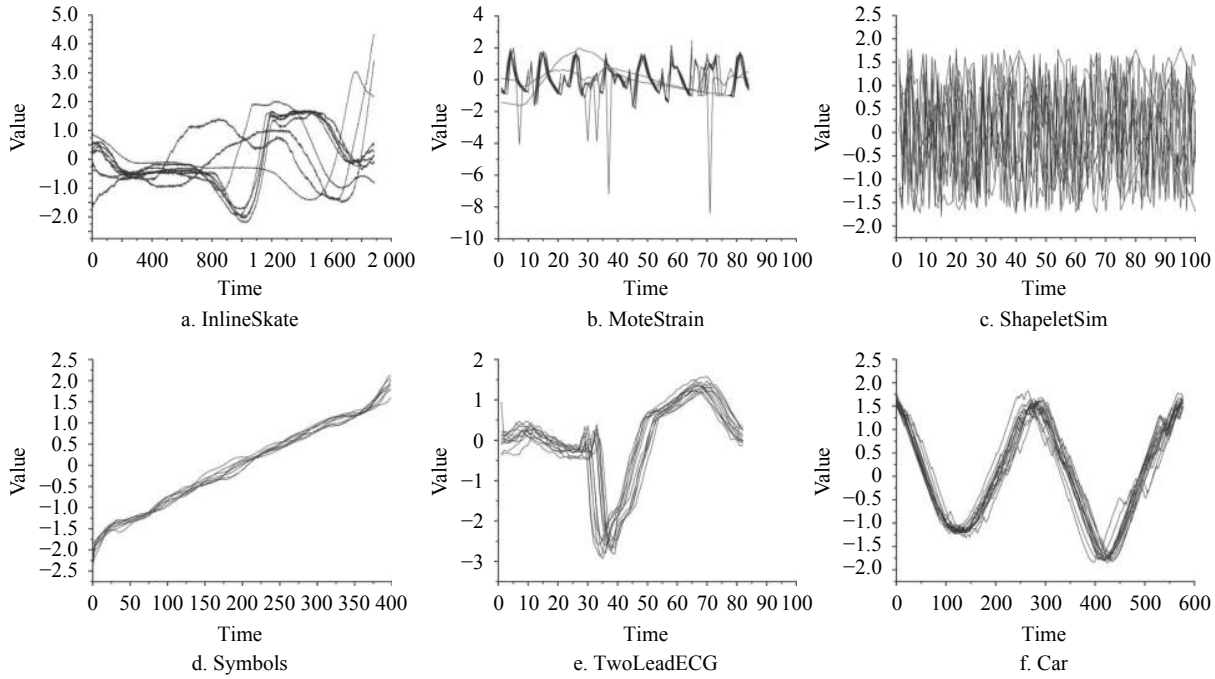


图 5　部分训练集中 label=1 的时序数据

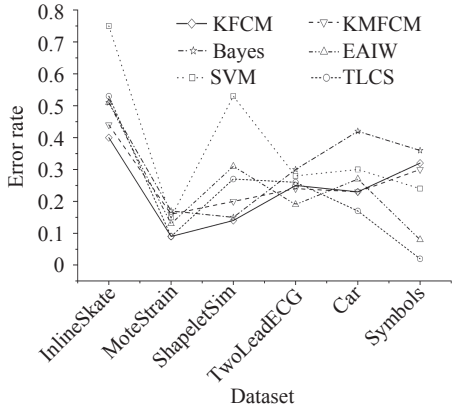


图 6　部分数据集错误率比较

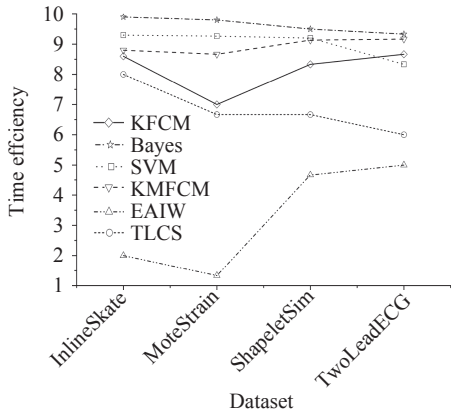


图 7　部分数据集时间效率比较

4　结 束 语

鉴于 *k*-shape 在时间序列数据聚类领域的优越性，本文提出了一种新的时间序列分类方法 KFCM。该方法首先利用 *k*-shape 对时间序列数据训练集中的各个类别包含的成员进行聚类，得到聚类中心群。然后，将聚类中心群作为模糊分类的初始聚类中心，根据隶属度最大原则确定测试时间序列数据的类别标签。通过实验验证，与传统时间序列分类方法相比，新分类方法具有以下优势：1) 通过使用 *k*-shape 聚类算法，提高了对具有位移和扭曲特征的时间序列数据集分类的适用性，在一定程度上弥补了传统聚类算法以欧氏距离作为相似性指标的不足；2) 新分类方法可以利用手肘法确定最佳聚类数，减小参数变化对最终分类结果的影响；3) 与传统分类方法相比，新分类方法能够实现更好的分类效果，具有一定的优越性。然而，该方法相较于传统分类方法 SVM 和 Bayes 有较高的时间复杂度，在大型数据集上应用效果不佳，这是未来需要进一步研究的工作。

参　考　文　献

[1] WU X D, KUMAR V, QUINLAN J R, et al. Top 10 algorithms in data mining[J]. Knowledge and Information Systems, 2008, 14(1): 1-37.

[2] 李海林, 万校基. 基于簇中心群的时间序列数据分类方法[J]. 电子科技大学学报, 2017, 46(3): 625-630.
LI H L, WAN X J. Classification for time series data based on center sequences of clusters[J]. Journal of University of Electronic Science and Technology of China, 2017, 46(3): 625-630.

[3] 王子一, 商琳. 基于子段距离计算的时间序列分类方法[J].

小型微型计算机系统, 2018, 39(7): 1387-1389.
WANG Z Y, SHANG L. New method of timeseries classification based on sub-sequence distancecomputation [J]. Journal of Chinese Computer Systems, 2018, 39(7): 1387-1389.

[4] 李海林, 梁叶, 王少春. 时间序列数据挖掘中的动态时间弯曲研究综述[J]. 控制与决策, 2018, 33(8): 1345-1353.
LI H L, LIANG Y, WANG S C. Reviewon dynamic time warping in time series data mining[J]. Control and Decision, 2018, 33(8): 1345-1353.

[5] 原继东, 王志海. 时间序列的表示与分类算法综述[J]. 计算机科学, 2015, 42(3): 1-7.
YUAN J D, WANG Z H. Review of time series represention and classification techniques[J]. Computer Science, 2015, 42(3): 1-7.

[6] VAPNIK V. SVM method of estimating density, conditional probability, and conditional density[C]//IEEE International Symposium on Circuits & Systems. [S.l.]: IEEE, 2000, 749-752.

[7] 石洪波, 王志海, 黄厚宽, 等. 一种限定性的双层贝叶斯分类模型[J]. 软件学报, 2004, 15(2): 193-199.
SHI H B, WANG Z H, HUANG H K, et al. A restricted double-level Bayesian classificationmodel[J]. Journal of Software, 2004, 15(2): 193-199.

[8] 李建平, 王兴伟, 马连博, 等. 基于区间的时间序列分类算法的研究[J]. 网络空间安全, 2019, 10(8): 84-101.
LI J P, WANG X W, MA L B, et al. Research on interval time series classification algorithm[J]. Cyberspace Security, 2019, 10(8): 84-101.

[9] 林钱洪, 王志海, 原继东, 等. 基于趋势信息的时间序列分类方法[J]. 中国科学技术大学学报, 2019, 49(2): 139-148.
LIN Q H, WANG Z H, YUAN J D, et al. Trend information for time series classification[J]. Journal of University of Science and Technology of China, 2019, 49(2): 139-148.

[10] 于剑. 论模糊 C 均值算法的模糊指标[J]. 计算机学报, 2003, 26(8): 969-973.
YU J. On the fuzziness index of the FCM algorithms[J]. Chinese Journal of Computers, 2003, 26(8): 969-973.

[11] 张新波. 两阶段模糊 C-均值聚类算法[J]. 电路与系统学报, 2005, 10(5): 118-120.
ZHANG X B. Two stage fuzzy C-means clustering algorithm[J]. Journal of Circuits and Systems, 2005, 10(5): 118-120.

[12] PAPARRIZOS J, GRAVANO L, SELLIS T K, et al. k-shape: Efficient and accurate clustering of time series[C]//Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data. Melbourne, Victoria: ACM, 2015: 1855-1870.

[13] 陈书文, 覃华, 苏一丹. 最优正则化参数的核 FCM 聚类算法[J]. 小型微型计算机系统, 2018, 39(7): 1538-1541.
CHEN S W, QIN H, SU Y D. Kernel FCM clustering algorithm based on optimal regularization parameters[J]. Journal of Chinese Computer Systems, 2018, 39(7): 1538-1541.

[14] CHEN Y P, KEOGH E, HU Bing, et al. The UCR time series classification archive[EB/OL]. [2019-11-12]. http://www.cs.ucr.edu/eamonn/time_series_data/.

[15] 王潇笛, 刘俊勇, 刘友波, 等. 采用自适应分段聚合近似的典型负荷曲线形态聚类算法[J]. 电力系统自动化, 2019, 43(1): 110-118.
WANG X D, LIU J Y, LIU Y B, et al. Typical load curve morphology clustering algorithm using adaptive segmentation aggregation approximation[J]. Automation of Electric Power Systems, 2019, 43(1): 110-118.

编 辑 叶 芳