# 基于簇中心群的时间序列数据分类方法

李海林，万校基

(1. 华侨大学信息管理系 福建 泉州 362021；2. 华侨大学现代应用统计与大数据研究中心 福建 厦门 361021)

【摘要】分类算法是时间序列数据挖掘中极为重要的任务和技术，该文提出一种基于簇中心群的时间序列数据分类方法。该方法根据时间序列训练数据集中的类别标签进行簇划分，利用近邻传播算法分别对每个簇进行中心代表点选择，构造出各代表点的代表对象集；然后借助基于动态时间弯曲的均值中心方法对各代表对象集实现中心群计算，结合改进后的$K$近邻算法实现时间序列数据的分类。数值实验结果表明，与传统方法相比，新方法具有更好的分类效果和计算性能。

**关 键 词** 近邻传播；分类算法；数据挖掘；动态时间弯曲；时间序列

**中图分类号** TP273 **文献标志码** A doi:10.3969/j.issn.1001-0548.2017.03.024

# Classification for Time Series Data Based on Center Sequences of Clusters

LI Hai-lin and WAN Xiao-ji

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

**Abstract** Classification algorithm is one of the important tasks and techniques in the field of time series data mining. A classification method for time series data based on center sequences of clusters is proposed in this paper. Time series in the training set are divided into several clusters according to their labels, and every cluster picks out the representation objects using affinity propagation clustering and constructs the representation subset. The barycenter averaging method based on dynamic time warping is used to calculate the center group in which the improved $K$ nearest neighbors method is executed for time series classification. The experimental results demonstrated that the new method, compared to the traditional method, has better classification quality and calculation performance.

**Key words** affinity propagation; classification algorithm; data mining; dynamic time warping; 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]。在时间序列数据挖掘领域中，特别是金融时间序列数据存在时间高维性，使得传统分类算法不能直接有效地对时间序列数据进行分类，有碍于金融数据市场分析。部分学者通过数据降维与特征表示方法将高维时间序列数据挖掘进行特征提取，再结合传统聚类或分类算法实现特征对象的数据分类[2-3]。然而，由于数据降维和特征表示在一定程度上会丢失部分重要数据信息，传统方法不能很好地对时间序列数据进行有效分类。有成果研究表明[4]，最近邻分类算法是时间序列数据分类最为有效的方法，它能较好地实现时间序列数据分类和预测。传统分类算法的分类质量和计算效率在一定程度上取决于前期数据处理中特征表示和相似性度量等方法的性能[5-6]。基于动态时间弯曲的最近邻分类方法是一种通过匹配异步形态相似性来对具有共同波动特征的时间序列数据进行聚类或分类，它能够提高最近邻方法的分类质量，但其平方阶的时间复杂度在一定程度上影响了其在高维时间序列数据挖掘中的应用效果[7]。

鉴于基于动态时间弯曲距离的最近邻算法在时间序列数据分类中重要性和有效性[8]，本文从分类质量和效率两个角度出发，提出一种基于簇中心群的时间序列分类算法。该方法利用近邻传播聚类算法对训练集中的每个簇进行代表点计算，并找到各

收稿日期: 2015 – 11 – 21；修回日期：2016 – 06 – 18
基金项目: 国家自然科学基金(61300139); 福建省社会科学规划项目(FJ2016B076); 福建省自然科学基金(2015J01581)
作者简介: 李海林(1982 – ), 男, 副教授, 博士, 主要从事数据挖掘与人工智能等方面的研究.

代表点所对应的被代表对象集，利用基于动态时间弯曲的均值中心来描述每个被代表对象集，最后结合改进后的 $K$ 近邻算法来讨论在不同 $K$ 值下的分类情况。数值实验结果与分析表明，新方法具有更好的时间序列数据分类质量和计算性能。

# 1 相关理论基础

## 1.1 动态时间弯曲

动态时间弯曲(dynamic time warping, DTW)是时间序列数据挖掘领域中用来进行相似性度量的一种经典方法，其能较好地对时间序列数据进行形态匹配，进而得到反映时间序列相似性的最小距离[8]。

**定义 1** DTW是按一定的规则从两条时间序列数据中寻找一条最优弯曲路径 $P=[p_1,p_2,\cdots,p_W]$，使得该弯曲路径对应元素之间的距离总和最小，即：

$$\mathrm{DTW}(X,Y)=\min_P\sum_{w=1}^{W}d(p_w) \tag{1}$$

式中，$d(p_w)=D(i,j)=d(x_i,y_j)$，表示最优弯曲路径 $P$ 中来自不同时间序列数据对应元素之间的距离，通常使用欧氏距离来度量元素之间的距离，即 $d(x_i,y_i)=(x_i-y_i)^2$。基于动态规划方法和距离矩阵可以求解获得一条满足最优情况的路径，使得该路径中最后一个元素的累积距离最小，即 $\mathrm{DTW}(X,Y)=\boldsymbol{R}(n,m)$，且有：

$$\boldsymbol{R}(i,j)=\boldsymbol{D}(i,j)+\min\begin{cases}\boldsymbol{R}(i-1,j)\\ \boldsymbol{R}(i-1,j-1)\\ \boldsymbol{R}(i,j-1)\end{cases} \tag{2}$$

DTW能够有效地匹配两条时间序列中具有相似性形态的数据点，且代价矩阵 $\boldsymbol{R}$ 记录了最优弯曲路径的方向和反映两条时间序列之间相似性的最小距离 $\boldsymbol{R}(n,m)$。由于需要通过累积代价矩阵 $\boldsymbol{R}$ 获得最优弯曲路径 $\boldsymbol{P}$，使得其计算时间复杂度为 $O(nm)$，不利于较长时间序列之间的距离度量。

## 1.2 近邻传播聚类

近邻传播(affinity propagation, AP)聚类[9]是一种基于近邻信息传播的聚类算法，与其他无监督的机器学习方法一样[10-11]，具有较高效率的分类效果。AP聚类目的是找出若干个最优代表点，使得其与所代表对象相似性之和最大。

AP聚类算法将所有数据对象视为聚类中心，为每个样本点建立与其他样本点的吸引程度信息，即相似性矩阵 $\boldsymbol{S}$，其中任意 $i$ 和 $j$，相似性矩阵中元素 $s(i,j)=-\|x_i-y_j\|^2$。另外，AP聚类算法涉及3个重要参数：偏向参数、代表程度及合适程度。

**定义 2** 偏向参数 $p(i)$ 表示数据点 $i$ 被选作聚类中心的倾向程度，初始可以被赋予一个先验值，由样本 $i$ 与其他样本之间的相似性的中位值来确定。

**定义 3** 代表程度 $r(i,k)$ 是指由样本点 $x_i$ 指向样本点 $x_k$，表示代表点 $x_k$ 积累的信息，用来说明 $x_k$ 作为 $x_i$ 的类代表点的程度。

**定义 4** 合适程度 $a(i,k)$ 是从样本点 $x_k$ 指向样本点 $x_i$，表示代表点 $x_i$ 积累的信息，用来表示 $x_i$ 选择 $x_k$ 作为代表点的合适程度。

$$r(i,k)\leftarrow s(i,k)-\arg_{k'\mathrm{s.t.}k'\neq k}\{a(i,k')+s(i,k')\} \tag{3}$$

$$a(i,k)\leftarrow\begin{cases}\min\{0,r(k,k)+\\ \quad\sum_{i'\mathrm{s.t.}i'\notin\{i,k\}}\max\{0,r(i',k)\}\} & i\neq k\\ \sum_{i'\mathrm{s.t.}i'\neq k}\max(0,r(i',k)) & i=k\end{cases} \tag{4}$$

在AP聚类算法中，通过代表程度和合适程度两个信息量的交替更新，计算所有数据点的代表程度 $r(i,k)$ 和合适程度 $a(i,k)$ 之和，取和值最大的 $x_{k_0}$ 作为 $x_i$ 的代表点，$k_0=\arg\max_k(a(i,k)+r(i,k))$。

AP聚类算法每次需要重复交替更新 $a(i,k)$ 和 $r(i,k)$，使其在不同替代次数下，数据集中被聚类所构成的代表点不同，直到达到指定又迭代次数或最终代表点被确定不变为止。

## 1.3 均值中心序列

均值中心序列(DTW barycenter averaging, DBA)[12]是一种基于DTW的时间序列中心序列，利用启发式规则来计算时间序列数据集的中心。其基本思想是，在数据集 $\boldsymbol{X}=\{\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_N\}$ 中，首先通过初始化中心序列 $\boldsymbol{C}=[c_1,c_2,\cdots,c_T]$，再利用DTW算法计算 $\boldsymbol{X}_i$ 与中心序列 $\boldsymbol{C}$ 的弯曲路径 $\boldsymbol{P}_i$；对于每个 $i$ 值，根据 $\boldsymbol{P}_i$ 值从 $\boldsymbol{X}_i$ 中选取与中心序列中数据点 $c_j$ 相匹配的数据点集合 $\boldsymbol{X}_i(j_{a_i}:j_{b_i})$；最后计算所有数据点 $\boldsymbol{X}_i(j_{a_i}:j_{b_i})$ $(i=1,2,\cdots,N)$ 的平均值作为更新后中心序列 $c_j$ 的值，即：

$$c_{j'}=\sum_{i=1}^{N}\sum_{k=j_{a_i}}^{j_{b_i}}\boldsymbol{X}_i(k)\Big/\sum_{i=1}^{N}j_{b_i}-j_{a_i}+1 \tag{5}$$

通过 $\boldsymbol{C}'$ 更新 $\boldsymbol{C}$，即 $\boldsymbol{C}\leftarrow\boldsymbol{C}'$，重新获得描述时间序列数据集 $\boldsymbol{X}$ 的均值中心序列 $\boldsymbol{C}$，直到连续两次替代中均值中心序列收敛不变为止。基于DTW的均值中心序列能够反映原始时间序列数据的形态变化。另外，DBA能够用不同长度的中心序列来描述

数据集中不等长时间序列的形态变化关系。

## 2 新分类方法

新分类方法首先通过构建训练簇中心群来描述每个簇中的对象特征，结合基于最近距离的近邻算法实现对象特征集的近邻分类，使其具有较好的分类质量和计算性能。

### 2.1 簇中心群

大数据时代，数据量呈现出爆炸式增长，若用单一中心代表点或均值中心(univariate center object, UCO)来描述超大型数据对象集，其对所有数据对象的特征描述力显得不足。因此，随着同类数据量的增长和数据特征的频繁演化，需要提出一种能够动态描述同类数据特征的代表对象群，使其能够更好地表达同类数据的特征。

**定义 5** 对于数据集 $\boldsymbol{A}$，若 $a_0=\mathrm{Rep}(\boldsymbol{A})$，则 $a_0$ 为数据集 $\boldsymbol{A}$ 的代表对象，Rep是一个求解代表对象的函数，它可以是均值、中位数或众数等函数。

**定义 6** 簇中心群是对同类数据集中的若干个代表对象的集合，使得被代表对象离代表对象的距离之和最小。形式化讲，对于同一簇中的数据对象集 $\boldsymbol{A}=[a_1,a_2,\cdots,a_M]$，该数据集被划分成 $K$ 个子集，$\boldsymbol{B}=\{\boldsymbol{B}_1,\boldsymbol{B}_2,\cdots,\boldsymbol{B}_K\}$，簇中心群 $\boldsymbol{C}=[c_1,c_2,\cdots,c_K]$，使得 $c_k=\mathrm{Rep}(\boldsymbol{B}_k)$，其中 $\boldsymbol{B}_i\in\boldsymbol{A}$，$\boldsymbol{B}_i\cap\boldsymbol{B}_j=\varnothing$ 且 $\bigcup_{i=1}^{K}\boldsymbol{B}_i=\boldsymbol{A}$。

簇中心群是对同一簇中具有较小差异的数据子集的代表对象的集合，与传统单一代表对象相比，其具有更好的数据特征表现力，可以减小代表对象与被代表对象的距离误差，即：

$$\sum_{k=1}^{K}\sum_{i=1}^{|\boldsymbol{B}_k|}\|b_{ki}-c_k\|^2\leqslant\sum_{j=1}^{M}\|a_j-a_0\|^2 \tag{6}$$

式中，$|\boldsymbol{B}_k|$ 是数据子集 $\boldsymbol{B}_k$ 的模，表示 $\boldsymbol{B}_k$ 中具有数据对象的个数；$b_{ki}$ 表示 $\boldsymbol{B}_k$ 数据子集中的第 $i$ 个数据对象，即有 $\boldsymbol{B}_k=[b_{k1},b_{k2},\cdots,b_{k|B_k|}]$。

为了更好地挑选簇中心群，本文提出一种基于近邻传播聚类的中心群选择方法(AP based center group, APCG)。其基本思想为，对同一类簇中所有对象集使用近邻传播聚类算法进行自动聚类，生成 $K$ 个子簇，获得每个子簇的代表对象，再结合DBA算法以对应的代表对象为初始中心序列计算每个子簇的均值中心序列，所有子簇产生的均值中心序列集合被视为簇中心群。其算法过程如下：

基于AP聚类的时序簇中心群方法：$\boldsymbol{C}=\mathrm{APCG}(\boldsymbol{A})$。

输入：同簇中时序数据集 $\boldsymbol{A}=[a_1,a_2,\cdots,a_M]$，其中 $a_i$ 表示某一条时间序列。

输出：簇中心群 $\boldsymbol{C}=[c_1,c_2,\cdots,c_K]$，$c_k$ 表示簇中的第 $k$ 个中心代表对象。

1) 根据AP聚类算法，将同一簇划分成 $K$ 个子簇和相应的代表对象，即 $[\boldsymbol{B},\boldsymbol{C}']=\mathrm{AP}(\boldsymbol{A})$，$\boldsymbol{B}$ 和 $\boldsymbol{C}'$ 分别表示被划分的子簇集合和代表对象集，即 $\boldsymbol{B}=\{\boldsymbol{B}_1,\boldsymbol{B}_2,\cdots,\boldsymbol{B}_K\}$ 和 $\boldsymbol{C}'=[c_1',c_2',\cdots,c_K']$，$c_K'$ 表示第 $K$ 个子簇的中心代表对象。

2) 以 $c_K'$ 为初始中心序列，利用DBA算法计算对应子簇 $\boldsymbol{B}_k$ 的中心序列，即 $c_k=\mathrm{DBA}(\boldsymbol{B}_k,c_K')$。

3) 重复步骤2)，计算所有子簇$\boldsymbol{B}$的均值中心序列，最终获得簇中心群 $\boldsymbol{C}=[c_1,c_2,\cdots,c_K]$。

通过AP聚类能够将同簇数据集进行自适应地划分成若干子类(记为 $K$ 类)，每个子类用DBA来表示对应时间序列子集的特征。

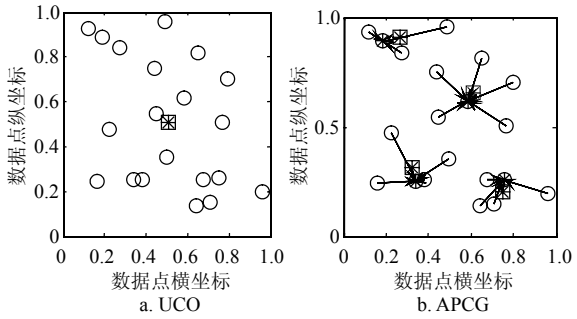



图1　基于单一均值和簇中心群的代表对象

如图1所示，方块和星号组合代表均值中心，子图b中圆圈和星号组合代表AP聚类算法产生的代表对象，箭头起始端表示被代表对象。若用单一均值代表点表示同一簇中的所有数据，其离差较大，代表中心对数据的代表能力较弱；相反，利用簇中心群中的对象对更相似的数据子集进行描述，将会产生具有较小的离差，说明具有较强的代表能力。

### 2.2 新 $K$ 近邻分类

在传统 $K$ 近邻分类算法中[13]，通过查找与被分类对象最相似或距离最近的前 $K$ 个数据对象，把被分类对象的类别归为这 $K$ 个对象中类别众数所对应的数据类标签。如图2所示，当近邻数为5时，被分类数据点 (0,0) 将被归为星号类。然而，从数据点之间的距离易知，被分类数据点与十字类2个数据点平均距离要小于与星号类3个数据点的平均距离，说明被分类数据点与十字类更相似。因此，传统最近邻算法不能很好地处理类似情况。特别地，当 $K$ 值较

小时，这种情况更容易发生。

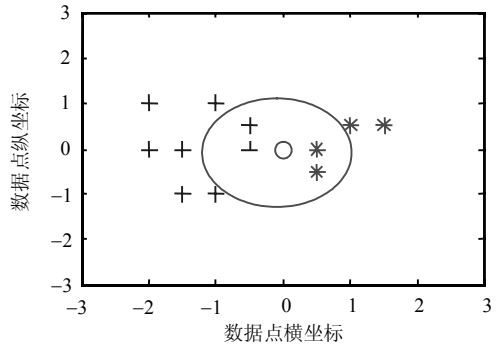



图2 $K$近邻分类算法分析

为了更好地使传统KNN算法适用于$K$值较小的时序分类，提出一种基于平均距离的$K$近邻分类方法(distance based KNN, DKNN)，其具体算法如下。

基于平均距离的 $K$ 近邻方法：$l=\mathrm{DKNN}(o,\boldsymbol{A},K)$。

输入：时序 $o$、训练集 $\boldsymbol{A}$ 和近邻数目 $K$。

输出：时序 $o$ 的类标签 $l$。

1) 利用DTW计算时序 $o$ 与 $\boldsymbol{A}$ 中所有时间序列 $a_j$ 的距离 $d_j\in\boldsymbol{D}$，即 $d_j=\mathrm{DTW}(o,a_j)$。

2) 根据距离向量 $\boldsymbol{D}$ 找出前 $K$ 个距离最小的数据对象集合 $\boldsymbol{S}$，根据它们的类别标签进行分组 $\boldsymbol{S}=\{\boldsymbol{S}_1,\boldsymbol{S}_2,\cdots,\boldsymbol{S}_w\}$，且标签记为 $\boldsymbol{L}=[l_1,l_2,\cdots,l_w]$，其中 $w$ 为 $K$ 个近邻对象的类数。

3) 计算每组 $\boldsymbol{S}_i$ 中数据对象与 $o$ 的平均距离 $\overline{d_i}$，记 $\overline{d_i}=\mathrm{averDist}(\boldsymbol{S}_i,\boldsymbol{D})$。

4) 将时序 $o$ 与平均距离最小的分组 $(k=\arg\min_i \overline{d_i})$对象视为同一类数据，并将该组类标签返回，即 $l=l_k$。

### 2.3 基于簇中心群的 $K$ 近邻分类方法

提出的基于簇中心群的 $K$ 近邻分类方法(KNN based on cluster center group, KNN2CG)利用APCG算法在训练集中对每个簇进行中心群计算，使得每个簇利用一个中心群来表示其总体特征。与此同时，将所有中心群成员对象视为新构建的训练数据集，对于测试集中的每个数据对象利用DKNN在新构建的训练集中实现分类。其具体算法如下。

基于簇中心群的 $K$ 近邻分类方法：$\boldsymbol{L}=\mathrm{KNN2CG}(\boldsymbol{A},\boldsymbol{B},K)$

输入：训练集 $\boldsymbol{A}$、测试集 $\boldsymbol{B}$ 和近邻数目 $K$。

输出：测试集 $\boldsymbol{B}$ 中成员类标签集合 $\boldsymbol{L}$。

1) 根据训练集 $\boldsymbol{A}=[a_1,a_2,\cdots,a_N]$ 中成员类标签划分成相应的簇，即 $\boldsymbol{A}=\{\boldsymbol{A}_1,\boldsymbol{A}_2,\cdots,\boldsymbol{A}_w\}$，且有 $\boldsymbol{A}=\bigcup_{i=1}^{w}\boldsymbol{A}_i$，其中 $w$ 为 $\boldsymbol{A}$ 中的类别数目。

2) 利用APCG对每个簇 $\boldsymbol{A}_i$ 计算其中心序列群，即 $C_i=\mathrm{APCG}(\boldsymbol{A}_i)$，进而获得簇中心群集合 $\boldsymbol{C}$ 且有 $\boldsymbol{C}=\bigcup_{i=1}^{w}\boldsymbol{C}_i$。

3) 对于测试集中的每个数据对象 $b_j$ 利用KNN2CG在簇中心群集合 $\boldsymbol{C}$ 中进行类标签预测，则有 $l_j=\mathrm{KNN2CG}(b_j,\boldsymbol{C},K)$。

4) 重复执行步骤3)，获得所有测试集中数据成员的预测类标签，即 $\boldsymbol{L}=[l_1,l_2,\cdots,l_M]$，其中 $M$ 表示测试集 $\boldsymbol{B}$ 中的成员数目。

新构建的特征训练集大小远小于原始训练集，使得DKNN能够快速有效地对时间序列进行分类。从时间效率角度来分析，KNN2CG方法的时间复杂度由训练集学习时间 $T_1$ 和测试集预测时间 $T_2$ 所决定，即：

$$T=T_1+T_2=O\left(N^2\left(t+\frac{m^2}{2}\right)+KMm^2N'\right) \tag{7}$$

由于新构建的训练集成员数量 $N'$ 远小于原始训练集成员数目 $N$，新方法预测时间将会远小于传统 $K$ 最近邻算法的时间，即 $KMm^2N'<KMm^2N$。因此，从时间复杂度分析可知，新方法具有更好的预测时间效率。

## 3 数值实验

### 3.1 实例分析

实例分析通过计算每个簇的中心群，用于验证簇中心群对相应簇成员的代表程度。从数据集Synthetic Control中随机选取30条时间序列数据，其也是金融市场中较为常见的股票波动现象，即存在6类趋势，分别为正常随机波动(No.: 1～5)、周期性波动(No.: 6～9)、上升波动趋势(No.: 10～13)、下降波动趋势(No.: 14～18)、向上跳跃势波动(No.: 19～23)和向下跳跃势波动(No.: 24～30)6种形态。

通过APCG方法，对6组时间序列数据中每组数据进行AP聚类划分，利用划分后的数据对象集进行均值中心序列计算，使得每组时间序列数据用中心群$\boldsymbol{C}$来反映每组时间序列数据的总体形态特征或者金融股票每个时间段所反映的群体波动趋势。如图3所示，每组数据被转化为若干个均值中心序列，均值中心序列的形态变化反映了原始时间序列数据的形态波动趋势。与此同时，根据每组数据的形态分

布情况，APCG对每组时间序列数据产生不同数量的均值中心序列。例如，APCG在前5组分别产生了2条均值中心序列，第6组却产生了3条均值序列。这说明APCG在训练数据集的学习过程中具有自适应性，同时也从体现了基于APCG的新方法的可靠性。

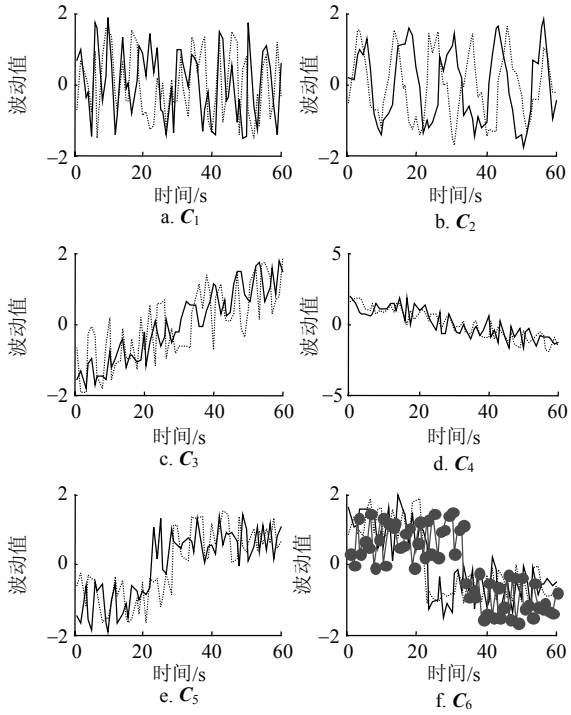



图3　训练集中各簇的中心群

### 3.2　分类实验

利用两种方法对15组UCI时间序列数据集[14]进行分类试验，具体实验数据信息如表1所示。通过统计测试集中数据成员对象预测标签的平均错误率来反映算法在对应时间序列数据集的分类质量，实验结果如表2和表3所示。

**表1　UCI时间序列数据集信息**

| 序号 | 数据集名称 | 类别数目 | 训练集大小 | 测试集大小 | 长度 |
|---|---|---|---|---|---|
| 1 | Adiac | 37 | 390 | 391 | 176 |
| 2 | Beef | 5 | 30 | 30 | 470 |
| 3 | CBF | 3 | 30 | 900 | 128 |
| 4 | ECG200 | 2 | 100 | 100 | 96 |
| 5 | Fish | 7 | 175 | 175 | 463 |
| 6 | FaceAll | 14 | 560 | 1 690 | 131 |
| 7 | Gun_Point | 2 | 50 | 150 | 150 |
| 8 | Lighting2 | 2 | 60 | 61 | 637 |
| 9 | Lighting7 | 7 | 70 | 73 | 319 |
| 10 | OSULeaf | 6 | 200 | 242 | 427 |
| 11 | OliveOil | 4 | 30 | 30 | 570 |
| 12 | SwedishLeaf | 15 | 500 | 625 | 128 |
| 13 | Trace | 4 | 100 | 100 | 275 |
| 14 | TwoPatters | 4 | 1 000 | 4 000 | 128 |
| 15 | Syn.Control | 6 | 300 | 300 | 60 |

**表2　KNN2CG方法的时间序列分类结果**

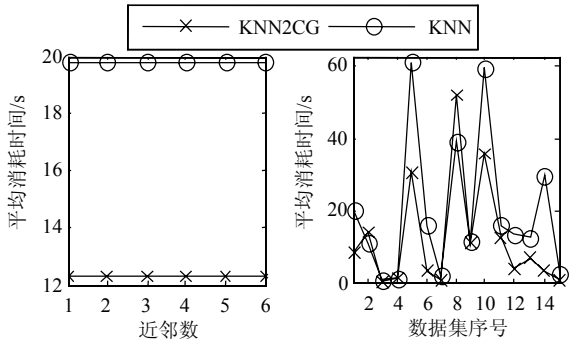
| 序号 | $K$ | | | | | | Aver |
|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | |
| 1 | 0.383 6 | 0.383 6 | 0.383 6 | 0.383 6 | 0.383 6 | 0.383 6 | 0.383 6 |
| 2 | 0.500 0 | 0.500 0 | 0.500 0 | 0.500 0 | 0.500 0 | 0.500 0 | 0.500 0 |
| 3 | 0.003 3 | 0.003 3 | 0.004 4 | 0.004 4 | 0.004 4 | 0.004 4 | 0.004 1 |
| 4 | 0.130 0 | 0.130 0 | 0.170 0 | 0.180 0 | 0.170 0 | 0.260 0 | 0.173 3 |
| 5 | 0.234 3 | 0.234 3 | 0.257 1 | 0.268 6 | 0.274 3 | 0.274 3 | 0.257 1 |
| 6 | 0.143 2 | 0.143 2 | 0.155 0 | 0.174 6 | 0.187 0 | 0.198 2 | 0.166 9 |
| 7 | 0.173 3 | 0.173 3 | 0.393 3 | 0.213 3 | 0.213 3 | 0.213 3 | 0.230 0 |
| 8 | 0.114 8 | 0.114 8 | 0.082 0 | 0.180 3 | 0.114 8 | 0.114 8 | 0.120 2 |
| 9 | 0.219 2 | 0.219 2 | 0.205 5 | 0.205 5 | 0.205 5 | 0.205 5 | 0.210 0 |
| 10 | 0.413 2 | 0.413 2 | 0.409 1 | 0.425 6 | 0.446 3 | 0.446 3 | 0.425 6 |
| 11 | 0.133 3 | 0.133 3 | 0.133 3 | 0.133 3 | 0.133 3 | 0.133 3 | 0.133 3 |
| 12 | 0.232 0 | 0.232 0 | 0.238 4 | 0.251 2 | 0.260 8 | 0.273 6 | 0.248 0 |
| 13 | 0.000 0 | 0.000 0 | 0.020 0 | 0.090 0 | 0.090 0 | 0.090 0 | 0.048 3 |
| 14 | 0.000 0 | 0.000 0 | 0.000 0 | 0.000 0 | 0.000 0 | 0.000 0 | 0.000 0 |
| 15 | 0.013 3 | 0.013 3 | 0.013 3 | 0.010 0 | 0.016 7 | 0.010 0 | 0.012 8 |
| Aver | 0.179 6 | 0.179 6 | 0.197 7 | 0.201 4 | 0.200 0 | 0.207 2 | — |

**表3　传统KNN方法的时间序列分类结果**

| 序号 | $K$ | | | | | | Aver |
|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | |
| 1 | 0.396 4 | 0.434 8 | 0.427 1 | 0.445 0 | 0.468 0 | 0.475 7 | 0.441 2 |
| 2 | 0.500 0 | 0.500 0 | 0.600 0 | 0.533 3 | 0.566 7 | 0.533 3 | 0.538 9 |
| 3 | 0.003 3 | 0.017 8 | 0.003 3 | 0.023 3 | 0.017 8 | 0.036 7 | 0.017 0 |
| 4 | 0.230 0 | 0.160 0 | 0.200 0 | 0.180 0 | 0.210 0 | 0.200 0 | 0.196 7 |
| 5 | 0.165 7 | 0.205 7 | 0.194 3 | 0.217 1 | 0.262 9 | 0.251 4 | 0.216 2 |
| 6 | 0.192 3 | 0.229 0 | 0.192 3 | 0.186 4 | 0.189 9 | 0.188 8 | 0.196 4 |
| 7 | 0.093 3 | 0.133 3 | 0.113 3 | 0.173 3 | 0.173 3 | 0.180 0 | 0.144 4 |
| 8 | 0.131 1 | 0.114 8 | 0.131 1 | 0.131 1 | 0.180 3 | 0.163 9 | 0.142 1 |
| 9 | 0.274 0 | 0.328 8 | 0.287 7 | 0.260 3 | 0.246 6 | 0.246 6 | 0.274 0 |
| 10 | 0.409 1 | 0.442 1 | 0.421 5 | 0.433 9 | 0.454 5 | 0.466 9 | 0.438 0 |
| 11 | 0.133 3 | 0.166 7 | 0.133 3 | 0.200 0 | 0.233 3 | 0.233 3 | 0.183 3 |
| 12 | 0.208 0 | 0.201 6 | 0.220 8 | 0.206 4 | 0.212 8 | 0.228 8 | 0.213 1 |
| 13 | 0.000 0 | 0.000 0 | 0.000 0 | 0.000 0 | 0.000 0 | 0.000 0 | 0.000 0 |
| 14 | 0.000 0 | 0.000 0 | 0.000 0 | 0.000 0 | 0.000 0 | 0.000 0 | 0.000 0 |
| 15 | 0.006 7 | 0.010 0 | 0.016 7 | 0.013 3 | 0.026 7 | 0.023 3 | 0.016 1 |
| Aver | 0.182 9 | 0.196 3 | 0.196 1 | 0.200 2 | 0.216 2 | 0.215 3 | — |

Aver列表示不同 $K$ 值下两种方法在对应时间序列数据集中的平均分类错误率，可以发现，本文提出的方法在大部分数据集中具有较小的平均错误率，说明新方法具有更好的分类质量。Aver行表示两种方法在不同数据集中对应同一个近邻数 $K$ 的平均分类错误率，对于大部分近邻数 $K$，本文提出的KNN2CG方法也具有较低的平均错误率。特别地，当 $K=1$ 时，两种方法成为了最近邻分类方法，而新方法KNN2CG具有比传统最近邻算法更好的分类质量。

针对每个数据集中的分类实验，记录在不同近邻数 $K$ 值的情况下两种方法所花费的时间，从不同近邻数和不同数据集的两个角度来观察两种方法的时间效率，实验结果如图4所示。




a. 近邻数不同的时间代价 b. 数据集不同的时间代价

图4 两种方法在不同数据集和$K$的平均消耗时间

图4a显示了两种方法的时间消耗量会随着近邻数 $K$ 值的增大而稍微变大，同时也说明了新方法KNN2CG在不同$K$值下的时间效率明显要优于传统KNN分类方法。图4b中的结果说明，在大部分数据集中，新方法的时间效率要优于KNN。相对于测试集来说，较小的训练集且较长的时间序列数据对象容易使KNN2CG获得较好的时间效率。

## 4 结 束 语

鉴于最近邻算法在时间序列分类研究中的重要性和优越性，提出了一种基于簇中心群的时间序列数据分类方法(KNN2CG)。通过近邻传播AP聚类对训练数据集中的每个簇进行子簇划分和代表对象选择，再以代表对象为初始化中心对象，利用DBA对每个子簇进行中心序列计算，进而构建训练簇中心群。同时，结合改进的 $K$ 最近邻方法，使得基于簇中心群的分类算法获得更好的分类效果和计算性能。新方法具有以下几点优势：1) 通过AP和DBA使得具有极为相似形态的时间序列数据子集被均值中心序列所描述，减少了新训练集中成员数量，提高了分类算法的计算性能。2) 中心群为每个簇提供了更为详细的总体特征描述，结合DTW使得均值中心序列能够更好地表达被描述对象的形态特征，有利于提高最近邻算法的分类质量。3) 利用平均距离来选取 $K$ 个近邻对象，克服了传统 $K$ 近邻方法限入局最优的问题。实验结果表明，与传统方法相比，新方法具有更好的分类质量和较高的计算效率。

本文研究工作还得到福建省高等学校新世纪优秀人才支持计划(Z1625112)和华侨大学中青年教师科研提升资助计划(ZQN-PY220)的资助，在此表示感谢。

## 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, 杨丽彬. 时间序列数据降维及特征表示新方法[J]. 控制与决策, 2013, 28(11): 1718-1722.
LI Hai-lin, YANG Li-bin. Method of dimensionality reduction and feature representation for time series[J]. Control and Decision, 2013, 28(11): 1718-1722.

[3] 李正欣, 郭建胜, 惠晓滨, 等. 基于共同主成分的多元时间序列降维方法[J]. 控制与决策, 2013, 28(4): 531-536.
LI Zheng-xin, GUO Jian-sheng, HUI Xiao-bin, etal. Dimension reduction method for multivariate time series based on common principal component[J]. Control and Decision, 2013, 28(4): 531-536.

[4] PETITJEAN F, FORESTIER G, NICHOLSON A, et al. Dynamic time warping averaging of time series allows faster and more accurate classification[C]//IEEE International Conference on Data Mining. Piscataway: IEEE, 2014: 470-479.

[5] 郭兴明, 袁志会, 丁晓蓉. 经验模式分解及关联维数在心音信号分类识别中的应用[J]. 电子科技大学学报, 2013, 42(6): 954-960.
GUO Xing-ming, YUAN Zhi-hui, DING Xiao-rong. Application of EMD and correlation dimension in classification and recognition of heart sound[J]. Journal of University of Electronic Science and Technology of China, 2013, 42(6): 954-960.

[6] KAYA H, GÜNDÜZ-ÖĞÜDÜCÜ. A distance based time series classification framework[J]. Information Systems, 2015, 51: 27-42.

[7] LI Hai-lin. Asynchronism-based principal component analysis for time series data mining[J]. Expert Systems with Applications, 2014, 41(6): 2842-2850.

[8] KEOGH E. Exact indexing of dynamic time warping[J]. Knowledge and Information Systems, 2005, 7(3): 358-386.

[9] FREY B J, DUECK D. Clustering by passing messages between data points[J]. Science, 2007, 315(5814): 972-976.

[10] 杨燕, 冯晨菲, 贾真. 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.
YANG Yan, FENG Chen-fei, JIA Zhen. A link-based fuzzy clustering ensemble[J]. Journal of University of Electronic Science and Technology of China, 2014, 43(6): 887-892.

[11] LIAO T W. Clustering of time series data survey[J]. Pattern Recognition, 2005, 38(11): 1857-1874.

[12] PETITJEAN F, KETTERLIN A , GANCARSKI P. A global averaging method for dynamic time warping, with applications to clustering[J]. Pattern Recognition, 2011, 44: 678-693.

[13] LEE Y, WEI C, CHENG T. Nearest-neighbor-based approach to time-series classification[J]. Decision Support Systems, 2012, 53(1): 207-217.

[14] KEOGH E, ZHU Q, HU B, et al. The UCR time series classification/clustering homepage[EB/OL]. [2015-06-08]. http://www.cs.ucr.edu/~eamonn/time_series_ data/.

编 辑 蒋 晓