第51卷第5期
2019年10月

南京航空航天大学学报
Journal of Nanjing University of Aeronautics & Astronautics

Vol. 51 No. 5
Oct. 2019

DOI:10.16356/j.1005-2615.2019.05.006

基于字典学习的混合采样数据分类方法

杨 倩[1] 于 洪[1] 李 劼[2] 谢永芳[3]

(1.重庆邮电大学计算智能重庆市重点实验室,重庆,400065;2.中南大学冶金与环境学院,长沙,410083;
3.中南大学信息科学与工程学院,长沙,410083)

摘要: 混合采样数据不仅仅具有不同采样频率数据之间特征集合不同,还有样本数量不一致等特点,传统的分类方法不能直接使用。因此,本文提出一种基于Fisher判别准则字典学习的混合采样数据分类方法以处理采样数据的分类任务。该模型巧妙借助处理多视图数据的分类思想,利用基于Fisher判别准则的字典学习方法,生成的结构化字典的每个原子与数据的类标签相关,同时采用Fisher判别准则使类内散度更小,类间散度更大来约束编码系数矩阵,从而大大提升分类性能。此外,本文针对混合采样数据的样本数量不一致特点,设计了混合采样数据判别分析模型的分类方案。最后实验结果验证了本文方法的有效性。

关键词: 字典学习;分类;Fisher判别;混合采样数据;多视图

中图分类号: TP391　　**文献标志码:** A　　**文章编号:** 1005-2615(2019)05-0618-07

Classification Method for Mixed Sampling Data Based on Dictionary Learning

YANG Qian[1], *YU Hong*[1], *LI Jie*[2], *XIE Yongfang*[3]

(1. Chongqing Key Laboratory of Computational Intelligence, Chongqing University of Posts and Telecommunications, Chongqing, 400065, China; 2. College of Metallurgy and Environment, Central South University, Changsha, 410083, China; 3. College of Information Science and Engineering, Central South University, Changsha, 410083, China)

Abstract: Mixed sampling data, whose different features are collected at different sampling frequencies, pervasively exists in the real world. Because the different sampling data set not only has different features, but also has different number of samples, traditional classification methods cannot be used directly. Therefore, this paper proposes a classification method for mixed sampling data based on Fisher discrimination dictionary learning to solve the classification problem of mixed sampling data. Inspired on some classification ideas for multi-view data, the proposed model compares multiple data collected at multiple sampling frequencies to multiple views of multi-view data, and designs a way to learn a sub-dictionary for each class of each view. A structured dictionary whose dictionary atoms have correspondence to the class labels is learned, so that the within-class scatter is less and the between-class scatter is bigger based on the Fisher discrimination criterion. In addition, this paper designs a specific classification scheme for the inconsistent sample size of mixed sampling data. Finally, experimental results demonstrate the effectiveness of the proposed model.

Key words: dictionary learning; classification; Fisher discrimination; mixed sampling data; multi-view

大数据时代已经来临,随着各种数据采集设备(如红外线摄像机和网络摄像机)或是不同的媒介(如文本、视频和音频)的出现和发展,数据的来源、形式越来越多样化[1]。在实际生活中,数据的采集频率会因为采集成本代价不同而高低有别,例如获取工业生产铝的电解槽数据时,电解质水平、铁含

基金项目: 国家自然科学基金(61876027, 61751312, 61533020)资助项目。
收稿日期: 2019-05-05;**修订日期:** 2019-07-03
通信作者: 于洪,女,教授,E-mail: yuhong@cqupt.edu.cn。

引用格式: 杨倩,于洪,李劼,等. 基于字典学习的混合采样数据分类方法[J]. 南京航空航天大学学报,2019,51(5):618-624. YANG Qian, YU Hong, LI Jie. Classification Method for Mixed Sampling Data Based on Dictionary Learning[J]. Journal of Nanjing University of Aeronautics & Astronautics,2019,51(5):618-624.

量和硅含量等数据需要专家进行实验采集，成本高、代价大，采样频率低，而电解槽工作电压、工作电流和针振等数据可以通过传感器实时上传，成本低，采样频率高。把这类描述同一对象，而来自不同采样频率、有不同特征集合的数据称为混合采样数据。传统的处理混合采样数据的一种方法是转换高频数据，使之与低频数据的采样频率相匹配[2]，即将混合采样数据处理成同低频数据，然而这种方式不可避免的存在高频数据信息丢失的问题。因此，本文方法不采用对混合采样数据做同频处理的方式，而是希望尽可能地利用原始数据达到提升混合采样数据分类性能的目的。

受多视图数据特点的启发，我们认为混合采样数据与多视图数据之间有一定的对应关系。混合采样数据的多个采样频率数据、不同采样频率下的不同特征集合就对应着多视图数据的多个视图数据、不同视图下的不同特征集合。因此，本文欲借鉴处理多视图数据的思想或方法，来解决混合采样数据分类问题。

多视图学习的重点是揭示不同视图的多个异构数据表的相关性[3]。近年来，字典学习[4,5]已经成为多视图数据分类的热点研究方法。Yang 等[6]引入 Fisher 判别准则，提出 Fisher 判别字典学习方法(Fisher discrimination dictionary learning, FDDL)，构建字典的列向量与类标签相对应的结构化字典，并验证了判别保真项和 Fisher 判别准则都有提升分类准确性的优势。Zhuang 等[7]利用样本类标签信息学习多模态判别字典和不同模态之间的映射函数以恢复缺失模态。基于 Hilbert - Schmidt 独立性准则，Gangeh 等[8]提出两个多视图字典学习技术，两个技术的不同之处在于：(1)对每个视图学习一个字典，并在学习字典的空间融合稀疏系数矩阵；(2)在所有视图的特征融合空间学习一个字典和相应的系数矩阵。Jing 等[9]分析不同视图对应的字典间原子向量的不相关性约束，提出不相关多视图判别字典学习方法(Uncorrelated multi-view discrimination dictionary learning, UMD2L)，学习多个视图的不相关判别字典。随后该研究小组 Wu 等[10]提出多视图低秩字典学习方法(Multi-view low-rank dictionary learning, MLDL)，为不同视图不同类的子字典添加低秩约束以适应噪声数据，同时增加不同视图的字典不冗余约束，使视图间字典相互独立。Wu 等[11]提出一种学习多视图间共享结构化字典方法，同时将视图间字典结构不相关性引入共享字典学习的过程，从而更有效地利用不同视图的互补信息。Wang 等[12]将分析字典学习引入多视图分类任务场景，应用边缘化目标函数学习策略来改进模型的分类性能。以往的多视图字典学习方法训练和测试阶段耗时严重，主要是因为采用 l_0 或 l_1 范数进行稀疏约束，所以 Wu 等[13]又提出了多视图综合与分析字典学习(Multi-view synthesis and analysis dictionaries learning, MSADL)，每个视图的判别字典包含一个结构化综合字典和一个分析字典，由综合字典获得每个类簇的样本重构信息，分析字典获得具判别信息的编码系数矩阵。字典学习逐渐成为大数据分析与处理中的重要话题与研究方向，但是，很少有研究将字典学习运用到混合采样数据分类任务中。

混合采样数据分类任务的重点是探索不同采样频率数据的相关性，极大限度地利用原始数据进行学习。因此，本文欲借鉴多视图字典学习方法，提出一种基于字典学习的混合采样数据分类方法，旨在学习多个采样频率数据的判别子字典，和以 Fisher 判别准则约束的更具判别信息的编码系数矩阵，以提升模型的分类性能。

1　预备知识

1.1　字典学习

给定数据样本 $\boldsymbol{A}=[a_1,a_2,\cdots,a_n]$，$\boldsymbol{A}\in\mathbf{R}^{n\times d}$，字典学习的目标任务是寻找字典 $\boldsymbol{D}=[d_1,d_2,\cdots,d_n]$，$\boldsymbol{D}\in\mathbf{R}^{n\times d}$，以及对应的系数矩阵 $\boldsymbol{X}=[x_1,x_2,\cdots,x_d]$，$\boldsymbol{X}\in\mathbf{R}^{d\times d}$，使得每个数据样本能更好地被字典重构，即 $a_i=\boldsymbol{D}x_i$，$i=1,2,\cdots,n$。字典学习解决以下优化问题

$$\{\hat{\boldsymbol{D}},\hat{\boldsymbol{X}}\}=\arg\min_{\boldsymbol{D},\boldsymbol{X}}\|\boldsymbol{A}-\boldsymbol{D}\boldsymbol{X}\|_{\mathrm{F}}^{2}+\lambda\|\boldsymbol{X}\|_{1}\quad(1)$$

显然式(1)的第 1 项是希望由系数矩阵 $\boldsymbol{X}$ 能很好地重构数据样本 $\boldsymbol{A}$，第 2 项则是希望系数矩阵 $\boldsymbol{X}$ 尽量稀疏，可以使用 KSVD 算法[14]或 MOD 算法[15]求解上述优化问题中字典和系数矩阵变量。

1.2　Fisher 判别字典学习

Fisher 判别字典学习(Fisher discrimination dictionary learning, FDDL)旨在学习一个结构化字典 $\boldsymbol{D}=[D_1,D_2,\cdots,D_C]$，$D_i$ 是每个类簇的字典，C 是样本的总类簇数。给定训练样本 $\boldsymbol{A}=[A_1,A_2,\cdots,A_C]$，$A_i$ 是第 i 类训练样本，定义 $\boldsymbol{X}$ 是样本矩阵 $\boldsymbol{A}$ 在字典 $\boldsymbol{D}$ 上的编码系数矩阵，$\boldsymbol{X}=[X_1,X_2,\cdots,X_C]$，$X_i$ 是第 i 类训练样本 A_i 在字典 $\boldsymbol{D}$ 上的编码。FDDL 的目标函数形式为

$$\min_{D,X} r(A_i,\boldsymbol{D},\boldsymbol{X}_i)+\lambda_1\|\boldsymbol{X}\|_1+\lambda_2(\mathrm{tr}(S_W(\boldsymbol{X})-S_B(\boldsymbol{X}))+\eta\|\boldsymbol{X}\|_F^2) \tag{2}$$

FDDL 目标函数的第 1 项是 1 个保真项 $r(A_i,\boldsymbol{D},\boldsymbol{X}_i)=\|A_i-\boldsymbol{D}\boldsymbol{X}_i\|_F^2+\|A_i-\boldsymbol{D}_iX_i^i\|_F^2+\sum_{\substack{j=1\\j\neq i}}^{C}\|\boldsymbol{D}_jX_i^j\|_F^2$，除了要求字典 $\boldsymbol{D}$ 能重构第 i 类训练样本 A_i，还要子字典 $\boldsymbol{D}_i$ 也能很好地重构 A_i，且其他子字典不能重构 A_i。

$$\begin{aligned}S_W(\boldsymbol{X})&=\sum_{i=1}^{C}\sum_{x_k\in X_i}(\boldsymbol{x}_k-\boldsymbol{m}_i)(\boldsymbol{x}_k-\boldsymbol{m}_i)^{\mathrm{T}}\\ S_B(\boldsymbol{X})&=\sum_{i=1}^{C}n_i(\boldsymbol{m}_i-\boldsymbol{m})(\boldsymbol{m}_i-\boldsymbol{m})^{\mathrm{T}}\end{aligned} \tag{3}$$

第 3 项类内散度 $S_W(\boldsymbol{X})$ 和类间散度 $S_B(\boldsymbol{X})$ 是为了约束编码系数矩阵，$\boldsymbol{m}_i$ 和 $\boldsymbol{m}$ 分别是编码系数矩阵 $\boldsymbol{X}_i$ 和 $\boldsymbol{X}$ 的均值向量，n_i 表示第 i 类训练样本 A_i 的样本数量。

1.3 不相关多视图判别字典学习

给定 M 个视图数据 $A_k(k=1,\cdots,M)$，不相关多视图判别字典学习(Uncorrelated multi-view discrimination dictionary learning, UMD^{2}L)旨在对每个视图学习视图间不相关的字典 $D_k(k=1,\cdots,M)$ 和相应的系数矩阵 $\boldsymbol{X}_k(k=1,\cdots,M)$。UMD^{2}L 的目标函数形式如

$$\begin{aligned}&\min_{\substack{D_1,\cdots,D_M\\X_1,\cdots,X_M}}\sum_{k=1}^{M}\sum_{i=1}^{C}q(A_k^i,D_k,X_k^i)+\lambda\sum_{k=1}^{M}\|\boldsymbol{X}_k\|_1\\ &\text{s.t. Corr}(D_k,D_l)=0,l\neq k\end{aligned} \tag{4}$$

式中：第 1 项是 1 个判别保真项，$q(A_k^i,D_k,\boldsymbol{X}_k^i)=\|A_k^i-D_k\boldsymbol{X}_k^i\|_F^2+\|A_k^i-D_k^i\boldsymbol{X}_k^{ii}\|_F^2+\sum_{\substack{j=1\\j\neq i}}^{C}\|D_k^j\boldsymbol{X}_k^{ij}\|_F^2$，要求字典 D_k 有能力重构 A_k。C 是样本的总类簇数，A_k^i 是 A_k 的第 i 类训练样本，D_k^i 是 D_k 的第 i 类子字典，$\boldsymbol{X}_k^i$ 定义为样本 A_k^i 在字典 D_k 上的编码系数矩阵，$\boldsymbol{X}_k^{ij}$ 定义为样本 A_k^i 在字典 D_k^j 上的编码系数矩阵。$\mathrm{Corr}(D_k,D_l)=0$ 表示不同视图间字典 D_k 和 D_l 的不相关约束，有 $\mathrm{Corr}(D_k,D_l)=\frac{\sum_{i=1}^{N_k}\sum_{j=1}^{N_l}\mathrm{Corr}(\boldsymbol{d}_k^i,\boldsymbol{d}_l^j)}{N_k\cdot N_l}$。$N_k$ 和 N_l 分别表示字典 D_k 和字典 D_l 的原子数。$\mathrm{Corr}(\boldsymbol{d}_k^i,\boldsymbol{d}_l^j)=\frac{(\boldsymbol{d}_k^i-\bar{d}_k^i\cdot\boldsymbol{I})^{\mathrm{T}}(\boldsymbol{d}_l^j-\bar{d}_l^j\cdot\boldsymbol{I})}{\|\boldsymbol{d}_k^i-\bar{d}_k^i\cdot\boldsymbol{I}\|\cdot\|\boldsymbol{d}_l^j-\bar{d}_l^j\cdot\boldsymbol{I}\|}$ 表示了 $\boldsymbol{d}_k^i$(字典 D_k 的第 i 个字典原子)和 $\boldsymbol{d}_l^j$(字典 D_l 的第 j 个字典原子)的相关性，$\bar{d}_k^i$ 和 $\bar{d}_l^i$ 分别代表两个原子的均值。

2 混合采样数据分类模型

本节将多视图分类的思想融入混合采样数据分类模型中，首先系统地总结了混合采样数据分类模型的基本框架，然后分别描述了目标函数的判别保真项和判别系数项，随后给出了模型的目标函数，最后提出了混合采样数据分类方案。

2.1 模型框架

图 1 系统地总结了所提模型的主要框架，把它分为两个阶段：字典学习阶段和样本分类阶段。在字典学习阶段，假设训练样本有 H 个采样频率不同的数据集，$A_k(k=1,\cdots,H)$，利用 1.2 节中判别保真项和 1.3 节中判别系数项两个原则，根据 2.4

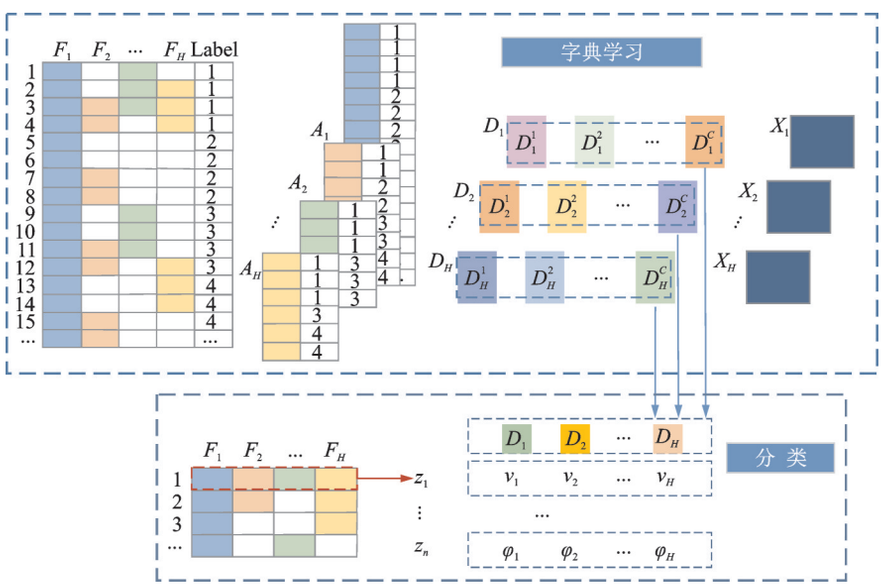

图 1 混合采样数据分类模型

Fig. 1 Classification model for mixed sampling data

节的目标函数学习每个采样频率数据集对应的每个类簇的子字典D_k^i和编码系数矩阵$\boldsymbol{X}_k^i$。在样本的分类阶段，与传统的测试样本不同，混合采样数据的测试样本会出现采样数据空缺现象，例如图1中分类阶段样本2在第3个采样频率数据集中没有对应值，样本3在第2个和第3个采样频率数据集中没有对应值。所以在2.5节，根据混合采样数据特点设计了混合采样数据的分类方案，利用对应采样频率数据的字典D_k对测试样本z进行稀疏编码，得到每个采样频率数据对应的编码向量$\boldsymbol{v}_k(k=1,\cdots,h)$，再利用子字典$D_k^i$和样本的编码向量判断样本与哪个类簇的重构误差最小，则表示样本属于该类簇。

2.2　判别保真项

假设训练数据有H个采样频率数据，定义为$A_k(k=1,\cdots,H)$。对于第k个采样频率数据，A_k^i表示第i类训练样本，D_k^i和D_k^j分别表示与第i类和第j类相关的类簇子字典。若$\boldsymbol{X}_k^i$定义为A_k^i在字典D_k上的编码系数矩阵，有$\boldsymbol{X}_k^i=[X_k^{i1},X_k^{i2},\cdots,X_k^{iC}]$，且$\boldsymbol{X}_k^{ij}$表示为样本$A_k^i$在子字典$D_k^j$上的编码系数矩阵，$A_k^i\approx D_kX_k^i=D_k^1X_k^{i1}+\cdots+D_k^iX_k^{ii}+\cdots+D_k^CX_k^{iC}$，$C$表示样本的类簇数。

目标函数的判别保真项应该遵循以下3个原则：(1)对于第k个采样频率数据，字典D_k应该具备重构样本A_k^i的能力，于是应该使误差项$\left\|A_k^i-D_kX_k^i\right\|_F^2$最小化。(2)与第$i$类相关的子字典$D_k^i$应该具备重构样本$A_k^i$的能力，于是应该使误差项$\left\|A_k^i-D_k^iX_k^{ii}\right\|_F^2$最小化。(3)与第$i$类相关的其他类子字典$D_k^j$不应具备重构样本$A_k^i$的能力，所以应该使$\left\|D_k^jX_k^{ij}\right\|_F^2$最小化，从而使第$i$类子字典$D_k^i$对样本$A_k^i$重构所占比重最大。判别保真项定义为

$$r(A_k^i,D_k,X_k^i)=\left\|A_k^i-D_kX_k^i\right\|_F^2+\left\|A_k^i-D_k^i\boldsymbol{X}_k^{ii}\right\|_F^2+\sum_{\substack{j=1\\j\neq i}}^{C}\left\|D_k^j\boldsymbol{X}_k^{ij}\right\|_F^2 \tag{5}$$

2.3　判别系数项

为了提升模型的分类性能，使字典D_k更具判别性，可以借助Fisher判别准则约束编码系数矩阵，最小化编码系数矩阵$\boldsymbol{X}_k$的类内散度$S_W(\boldsymbol{X}_k)$，最大化类间散度$S_B(\boldsymbol{X}_k)$，其定义如下

$$\begin{aligned}S_W(\boldsymbol{X}_k)&=\sum_{i=1}^{C}\sum_{\boldsymbol{x}_w\in X_k^i}(\boldsymbol{x}_w-\boldsymbol{m}_k^i)(\boldsymbol{x}_w-\boldsymbol{m}_k^i)^{\mathrm{T}}\\S_B(\boldsymbol{X}_k)&=\sum_{i=1}^{C}n_k^i(\boldsymbol{m}_k^i-\boldsymbol{m}_k)(\boldsymbol{m}_k^i-\boldsymbol{m}_k)^{\mathrm{T}}\end{aligned} \tag{6}$$

式中：$\boldsymbol{m}_k^i$和$\boldsymbol{m}_k$分别表示第k个采样频率数据第i类编码系数矩阵$\boldsymbol{X}_k^i$和$\boldsymbol{X}_k$的均值向量，n_k^i是样本A_k^i的样本数量。判别系数项要求类内散度$S_W(\boldsymbol{X}_k)$最小化，类间散度$S_B(\boldsymbol{X}_k)$最大化，很显然，如果直接定义$f(\boldsymbol{X}_k)$为$\mathrm{tr}(S_W(\boldsymbol{X}_k))-\mathrm{tr}(S_B(\boldsymbol{X}_k))$，则$f(\boldsymbol{X}_k)$是一个非凸且不稳定的函数，所以在判别系数项后增加一个弹性项$\left\|\boldsymbol{X}_k\right\|_F^2$使其满足凸优化条件，文献[6]已证明判别系数项的凸性。判别系数项定义为

$$f(\boldsymbol{X}_k)=\mathrm{tr}(S_W(\boldsymbol{X}_k))-\mathrm{tr}(S_B(\boldsymbol{X}_k))+\eta\left\|\boldsymbol{X}_k\right\|_F^2 \tag{7}$$

2.4　目标函数

结合判别保真项式(5)和判别系数项式(7)，混合采样数据判别分析模型的目标函数形式化定义为

$$\begin{aligned}&J(D_1,\cdots,D_H,X_1,\cdots,X_H)=\\&\underset{\substack{D_1,\cdots,D_H\\X_1,\cdots,X_H}}{\operatorname{argmin}}\sum_{k=1}^{H}\sum_{i=1}^{C}\mathrm{r}(A_k^i,\boldsymbol{D}_k,\boldsymbol{X}_k^i)+\\&\lambda_1\sum_{k=1}^{H}\left\|\boldsymbol{X}_k\right\|_1+\lambda_2\sum_{k=1}^{H}f(\boldsymbol{X}_k)\end{aligned} \tag{8}$$

2.5　混合采样数据分类方案

当混合采样数据对应各采样频率数据的类簇子字典D_k^i学习完成后，模型进入分类阶段。在文献[16]中，稀疏表示分类法(Sparse representation based classification, SRC)用原始各类簇样本作为结构化子字典，来编码测试样本，如果哪个类簇子字典对样本的重构误差越小，则表明样本属于该类簇。借鉴SRC分类方法和多视图分类思想，考虑到混合采样数据与多视图数据的不同之处，在于不同采样频率对应的数据集样本数量不一致，所以在测试阶段每个测试样本不一定涵盖所有采样频率对应数据，即测试样本$z=\{\boldsymbol{z}_1,\boldsymbol{z}_2,\cdots,\boldsymbol{z}_h\}$，$h\leqslant H$。那么$\boldsymbol{z}_k$在字典$D_k$上的编码向量可以通过下式求解

$$\begin{aligned}(\hat{v}_1,\hat{v}_2,\cdots,\hat{v}_h)=\arg\min_{v_1,\cdots,v_h}\{\sum_{k=1}^{h}(\left\|\boldsymbol{z}_k-\boldsymbol{D}_k\boldsymbol{v}_k\right\|_2^2+\\\beta\left\|\boldsymbol{v}_k\right\|_1)\}\end{aligned} \tag{9}$$

式中：$\boldsymbol{v}_k(k=1,\cdots,h)$是样本$\boldsymbol{z}_k$在字典$D_k$上的编码向量。定义$\hat{\boldsymbol{v}}_k=[\hat{v}_k^1;\hat{v}_k^2;\cdots;\hat{v}_k^C]$，$\hat{v}_k^1$对应样本$\boldsymbol{z}_k$在字典$D_k^i$上的编码向量。当测试样本对不同采样频率数据上字典进行的编码系数向量$\{\hat{v}_1,\cdots,\hat{v}_h\}$产生，将其应用到如下分类方案中，对给定测试样本$\boldsymbol{z}_k$，样本对每个类的重构误差计算为

$$e_i=\sum_{k=1}^{h}\left\|\boldsymbol{z}_k-D_k^i\hat{v}_k^i\right\|_2^2+w\left\|\hat{\boldsymbol{v}}_k-m_k^i\right\|_2 \tag{10}$$

$$\mathrm{identity}(z)=\arg\min_i\{e_i\} \tag{11}$$

式(10)第1项为样本$\boldsymbol{z}_k$由第i个类簇子字典重构的误差项，第2项为编码系数向量$\hat{\boldsymbol{v}}_k$与训练系数均值向量$\boldsymbol{m}_k^i$的距离。最后根据式(11)判断测试样本属

于哪一类。

3 优化目标函数

混合采样数据分类模型的目标函数中的变量可以在一个迭代过程中交替优化，可以使用以下优化策略进行每次迭代：(1)当更新第k($k=1,2,\cdots,H$)个采样频率数据对应的变量时，其他采样频率数据的对应变量固定。(2)对于第k个采样频率数据，$\boldsymbol{X}_k$和$\boldsymbol{D}_k$交替更新。

3.1 更新编码系数矩阵$\boldsymbol{X}_k$

假设$\boldsymbol{D}_k$固定，编码系数矩阵$\boldsymbol{X}_k=[X_k^1,X_k^2,\cdots,X_k^C]$，在遍历每个类计算$X_k^i$时，$X_k^j(j\neq i)$是固定的，目标函数式(8)可以简化为

$$J(X_k^i)=\arg\min_{X_k^i} r(A_k^i,\boldsymbol{D}_k,\boldsymbol{X}_k^i)+\lambda_1\left\|\boldsymbol{X}_k^i\right\|_1+\lambda_2 f(\boldsymbol{X}_k^i) \tag{12}$$

且有$f(X_k^i)=\left\|\boldsymbol{X}_k^i-\boldsymbol{M}_k^i\right\|_{\mathrm{F}}^2-\sum_{i=1}^{C}\left\|\boldsymbol{M}_k^i-\boldsymbol{M}_k\right\|_{\mathrm{F}}^2+\gamma\left\|\boldsymbol{X}_k^i\right\|_F^2$，其中$\boldsymbol{M}_k^i$和$\boldsymbol{M}_k$是两个均值向量矩阵(以$n_k^i$个$\boldsymbol{m}_k^i$或$\boldsymbol{m}_k$为列向量组成)。文献[6]已证明当$\gamma>1-n_k^i/n_k$时，$f(\boldsymbol{X}_k^i)$关于$\boldsymbol{X}_k^i$满足严格凸性，$n_k^i$和$n_k$分别是第$k$个采样频率数据属于第$i$类和所有类的样本数量。为了使$f(\boldsymbol{X}_k^i)$即满足凸性又保证足够的判别性，设定$\gamma=1$，所以式(12)中除了稀疏项$\left\|\boldsymbol{X}_k^i\right\|_1$是可微的，可以使用迭代投影算法[17](Iterative projection method, IPM)求解式(12)。迭代步长定义为δ，最终的求解依赖步长的取值，文献[17]中描述了步长及软阈值因子的公式。IPM算法见算法1。

算法1 混合采样数据系数矩阵$\boldsymbol{X}_k^i$更新算法

输入：δ和τ，$\delta>0$，$\gamma>0$

输出：$\boldsymbol{X}_k^i=\boldsymbol{X}_k^{i(h)}$

① 初始化：系数矩阵$\boldsymbol{X}_k^{i(1)}=0$，迭代次数$h=1$。

② 判断目标函数是否收敛或达到最大迭代次数，否则执行第3步，是执行第5步。

③ 增加迭代次数$h=h+1$。

④ 更新编码系数矩阵$\boldsymbol{X}_k^i$

$$\boldsymbol{X}_k^{i(h)}=S_{\tau/\delta}\left(\boldsymbol{X}_k^{i(h-1)}-\frac{1}{2\delta}\nabla Q(\boldsymbol{X}_k^{i(h-1)})\right)$$

其中：$\nabla Q(\boldsymbol{X}_k^{i(h-1)})$为$Q(\boldsymbol{X}_k^i)$在$\tilde{\boldsymbol{X}}_k^{i(h-1)}$处的导数，$S_{\tau/\delta}(\cdot)$是一个软阈值因子公式。

⑤ 输出编码系数矩阵$\boldsymbol{X}_k^i$，算法结束。

3.2 更新字典矩阵$\boldsymbol{D}_k$

当编码系数矩阵$\boldsymbol{X}_k$更新完，不同采样频率数据的子字典D_k^i逐类更新。在更新D_k^i时，所有$D_k^j(j\neq i)$固定，因此式(8)可以简化为

$$J(D_k^i)=\arg\min_{D_k^i,X_k^{ii}}\left\|A_k^i-D_k^iX_k^{ii}-\sum_{\substack{j=1\\j\neq i}}^{C}D_k^jX_k^{ij}\right\|_{\mathrm{F}}^2+\left\|A_k^i-D_k^iX_k^{ii}\right\|_F^2+\sum_{\substack{j=1\\j\neq i}}^{C}\left\|D_k^iX_k^{ij}\right\|_{\mathrm{F}}^2 \tag{13}$$

式(13)可以先将矩阵优化问题转换成向量优化按列更新字典，具体可以参见文献[18]中MFL算法。混合采样数据分类算法的整个迭代优化过程见算法2。

算法2 混合采样数据分类优化算法

输入：混合采样数据$A_k(k=1,\cdots,H)$，标签$L_k(k=1,2,\cdots,H)$

输出：字典矩阵$\{D_1,D_2,\cdots,D_H\}$和$\{X_1,X_2,\cdots,X_H\}$

① 用PCA方法初始化每个采样频率数据对应子字典D_k^i

$$D_k^i\leftarrow\mathrm{pca}(A_k^i)$$

② 判断目标函数是否收敛或达到最大迭代次数，否则执行第3步，是执行第5步。

③ 固定字典矩阵$\boldsymbol{D}_k$，对每个采样频率数据用算法1逐类更新编码系数矩阵X_k^i

$$X_k\leftarrow\mathrm{update}\,X^i(X_k^i)$$

④ 固定编码系数矩阵X_k，对每个采样频率数据用MFL算法[20]逐列更新字典矩阵$\boldsymbol{D}_k^i$

$$\boldsymbol{D}_k\leftarrow\mathrm{update}D^i(D_k^i)$$

⑤ 输出字典矩阵$\{D_1,D_2,\cdots,D_H\}$和$\{X_1,X_2,\cdots,X_H\}$，算法结束。

4 实验分析

本文算法采用Python语言并在Python 3.5解释器上进行实现，所有实验都在内存为8 GB RAM，CPU频率为2.70 GHz计算机上运行。为了便于实验，在UCI的一些真实多视图数据集上验证了本文提出的方法，表1给出了5个多视图数据集的基本信息。

WebKB数据集(http://membres-lig.imag.fr/grimal/data.html)包含了来自4个大学的网页信息，每个网页可以被内容和链接信息等4个视图描述。由于4个子数据集在内容上较为相似，选择Texas数据集上的3个视图进行实验。SensIT数据集(https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/multiclass.html)包含了用于划分3种车辆的2个传感器视图数据，每个传感器视图数据有50

个特征集合。Digit数据集(https://archive.ics.uci.edu/ml/datasets.html)描述手写数字0到9的特征信息，有10个类，2 000个样本，6个视图，选择以下3个视图进行后续实验：76维字符形状的Fourier系数特征，64维Karhunen-Love系数特征和47维Zernike moments特征。Cora数据集(http://membres-lig.imag.fr/grimal/data.html)描述了2 708个科学出版物文档，选择包含1 433个词特征的文档视图和含2 708个特征的文档链接视图。Citeseer数据集(http://membres-lig.imag.fr/grimal/data.html)描述了3 312个科学出版物文档，选择含3 703个词特征的文档视图和含3 312个特征的文档链接视图。

表1　UCI数据集

Tab. 1　The UCI data sets

Data set	Object	Dimension	View	Cluster
WebKB	187	{187, 187, 187}	3	5
SensIT	300	{50, 50}	2	3
Digit	2 000	{76, 64, 47}	3	10
Cora	2 708	{1 433, 2 708}	2	7
Citeseer	3 312	{3 703, 3 312}	2	6

本节实验主要包含两个部分：(1)实验数据集的各个视图以统一采样频率获取，以SensIT为例，利用本文所提出的方法，对SensIT的两个视图进行字典学习后分类，与将两个视图数据进行拼接后运用KNN分类算法进行对比。由图2实验结果看出，在WebKB，SensIT，Cora和Citeseer四个数据集上，本文提出模型的分类准确率比暴力拼接处理后的数据在KNN算法上的分类准确率略高，证明此类问题以拼接手段处理混合采样数据在某些场景是不合理的。

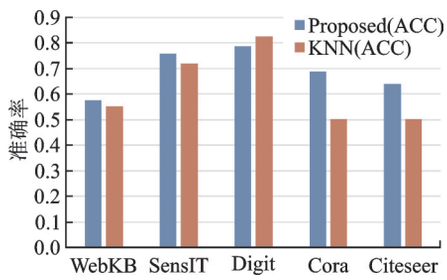

图2　实验结果1

Fig. 2　Experimental results 1

(2)实验在SensIT和Digit数据集上，对SensIT的每个视图数据划分200个样本作为训练集，100个样本作为测试集；Digit的每个视图数据划分1 500个样本为训练集，500个样本为测试集。再以不同程度的采样频率对数据的每个视图进行数据采样生成混合采样数据，与对混合采样数据进行均值或中位数填充后在改进的处理多个数据表的FDDL算法进行对比。由表2看出，在某些场景下对混合采样数据进行均值或中位数填充等手段是不合理的，从而验证了混合采样数据分类算法的有效性。

表2　实验结果2

Tab. 2　Experimental result 2

Index	Data set	Object	Sampling rate/%	Proposed	Improved FDDL[4]
ACC	SensIT	300	{100, 50}	0.63	0.46
			{100, 40}	0.64	0.57
	Digit	2 000	{100, 80, 60}	0.474	0.492
			{100, 70, 50}	0.512	0.446
			{100, 60, 40}	0.554	0.488
F1-score	SensIT	300	{100, 50}	0.623	0.449
			{100, 40}	0.632	0.514
	Digit	2 000	{100, 80, 60}	0.498	0.511
			{100, 70, 50}	0.535	0.453
			{100, 60, 40}	0.571	0.512

5　结　　论

在实际生活中，数据的采集频率会因为数据采集成本代价不同而高低有别，而采集成本高代价大的数据数量有限但是又相对重要，如何极大限度的利用原始数据，将多个不同采样频率数据充分运用成为一个研究难点。因此，本文将混合采样数据特有的不同采样频率、含不同特征集合的数据特点与多视图数据的多个视图、多个特征集合相类比，巧妙地借助多视图数据分类任务的思想，提出了一种基于字典学习的混合采样数据判别分析模型，该模型与多视图数据分类任务类似，但是又有其特有的混合采样频率的数据特点。本文方法在UCI的真实数据集上得到了验证，实验结果表明本文方法在

处理混合采样数据分类问题具有一定的效果。但是该方法在目标函数的字典迭代更新过程中采用逐列更新,计算量较大,时间复杂度较高,还需要进一步完善,以便于更好地处理样本数和特征维数较大的混合采样数据分类问题。

参考文献:

[1] NAKAMURA E F, LOUREIRO A A F, FRERY A C. Information fusion for wireless sensor networks: Methods, models, and classifications[J]. ACM Computing Surveys (CSUR), 2007, 39(3): 9.

[2] ARMESTO M T, ENGEMANN K M, OWYANG M T. Forecasting with mixed frequencies[J]. Federal Reserve Bank of St. Louis Review, 2010, 92(6): 521-536.

[3] ZHAO Jing, XIE Xijiong, XU Xin, et al. Multi-view learning overview: Recent progress and new challenges[J]. Information Fusion, 2017, 38: 43-54.

[4] TOSIC I, FROSSARD P. Dictionary learning[J]. IEEE Signal Processing Magazine, 2011, 28(2): 27-38.

[5] ABDI A, RAHMATI M, EBADZADEH M M. Dictionary learning enhancement framework: Learning a non-linear mapping model to enhance discriminative dictionary learning methods[J]. Neurocomputing, 2019, 357: 135-150.

[6] YANG Meng, ZHANG Lei, FENG Xiangchu, et al. Fisher discrimination dictionary learning for sparse representation [C]//International Conference on Computer Vision. N J: IEEE, 2011: 543-550.

[7] ZHUANG Yueting, WANG Yanfei, WU Fei, et al. Supervised coupled dictionary learning with group structures for multi-modal retrieval [C]// Twenty-Seventh AAAI Conference on Artificial Intelligence. CA: AAAI, 2013: 1070-1076.

[8] GANGEH M J, FEWZEE P, GHODSI A, et al. Multiview supervised dictionary learning in speech emotion recognition[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2014, 22(6): 1056-1068.

[9] JING Xiaoyuan, HU Ruimin, WU Fei, et al. Uncorrelated multi-view discrimination dictionary learning for recognition[C]//Twenty-Eighth AAAI Conference on Artificial Intelligence. CA: AAAI, 2014: 2787-2795.

[10] WU Fei, JING Xiaoyuan, YOU Xinge, et al. Multi-view low-rank dictionary learning for image classification[J]. Pattern Recognition, 2016, 50: 143-154.

[11] WU Fei, JING Xiaoyuan, YUE Dong. Multi-view discriminant dictionary learning via learning view-specific and shared structured dictionaries for image classification[J]. Neural Processing Letters, 2017, 45(2): 649-666.

[12] WANG Qianyu, GUO Yanqing, WANG Jiujun, et al. Multi-view analysis dictionary learning for image classification[J]. IEEE Access, 2018, 6: 20174-20183.

[13] WU Fei, DONG Xiwei, HAN Lu, et al. Multi-view synthesis and analysis dictionaries learning for classification[J]. IEICE Transactions on Information and Systems, 2019, 102(3): 659-662.

[14] AHARON M, ELAD M, BRUCKSTEIN A. K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation[J]. IEEE Transactions on Signal Processing, 2006, 54(11): 4311.

[15] ENGAN K, AASE S O, HUSOY J H. Method of optimal directions for frame design[C]//International Conference on Acoustics, Speech, and Signal Processing. N J: IEEE, 1999: 2443-2446.

[16] WRIGHT J, YANG A Y, GANESH A, et al. Robust face recognition via sparse representation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 31(2): 210-227.

[17] ROSASCO L, VERRI A, SANTORO M, et al. Iterative projection methods for structured sparsity regularization:MIT-CSAIL-TR-2009-050, CBCL-282[R].[S.l]: MIT, 2009:18-47.

[18] YANG Meng, ZHANG Lei, YANG Jian, et al. Metaface learning for sparse representation based face recognition[C]//International Conference on Image Processing. N J: IEEE, 2010: 1601-1604.

(编辑:刘彦东)