引用格式:段立娟, 郭亚静, 解晨瑶, 等. 基于特征聚类和等距映射的无监督特征选择算法[J]. 北京工业大学学报, 2024, 50(3): 325-332.

DUAN L J, GUO Y J, XIE C Y, et al. Unsupervised feature selection algorithm based on feature clustering and isometric mapping[J]. Journal of Beijing University of Technology, 2024, 50(3): 325-332. (in Chinese)

基于特征聚类和等距映射的无监督特征选择算法

段立娟[1,2,3], 郭亚静[1,2,3], 解晨瑶[1,2,3], 张文博[1,2,3]

(1. 北京工业大学信息学部, 北京 100124; 2. 可信计算北京市重点实验室, 北京 100124;
3. 信息安全等级保护关键技术国家工程实验室, 北京 100124)

摘 要: 为了提高无标签场景下特征选择的准确率和稳定性,提出一种基于特征聚类和等距映射的无监督特征选择算法。特征聚类将相似性较高的特征聚成一类,然后结合等距映射和稀疏系数矩阵定义新的特征得分计量函数。该函数对各特征簇中的特征进行打分,选择出每个类簇中得分最高的代表特征,构成特征子集。在14个广泛应用的数据集上的实验结果表明:本文所提算法能够选择出具有强分类能力的特征,且算法具有很强的泛化性。

关键词: 特征选择方法; 多源数据集; 高维特征; 无标签场景; 特征聚类; 等距映射

中图分类号: U461; TP308 **文献标志码**: A **文章编号**: 0254-0037(2024)03-0325-08

doi: 10.11936/bjutxb2022050016

Unsupervised Feature Selection Algorithm Based on Feature Clustering and Isometric Mapping

DUAN Lijuan[1,2,3], GUO Yajing[1,2,3], XIE Chenyao[1,2,3], ZHANG Wenbo[1,2,3]

(1. Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China;
2. Beijing Key Laboratory of Trusted Computing, Beijing 100124, China;
3. National Engineering Laboratory for Critical Technologies of Information Security Classified Protection, Beijing 100124, China)

Abstract: To improve the accuracy and stability of feature selection in label-free scenarios, an unsupervised feature selection algorithm based on feature clustering and isometric mapping was proposed. Feature clustering clustered features with high similarity into one class, and a new feature score measurement function was defined by combining isometric mapping and sparse coefficient matrix. This function scored the features in each feature cluster and selected the representative features with the highest scores in each class cluster to form a feature subset. Experimental results on fourteen widely used datasets show that the proposed algorithm can select features with strong classification ability and the algorithm is highly generalizable.

Key words: feature selection method; multi source data set; high dimensional features; unlabeled scene; feature clustering; isometric mapping

收稿日期: 2022-05-27; 修回日期: 2022-07-05

基金项目: 国家自然科学基金资助项目(62176009,62106065); 北京市教委科技计划重点资助项目(KZ201910005008)

作者简介: 段立娟(1973—), 女, 教授, 博士研究生导师, 主要从事人工智能、信息安全方面的研究, E-mail: ljduan@bjut.edu.cn

计算机技术的飞速发展使人们生活在一个万物互联的社会环境中,由于数据来源的多元化,大量高维数据分析已经成为很多领域的重要问题,比如计算机视觉[1-2]、数据挖掘[3-5]和模式识别[6]。

这些海量高维数据中存在着大量的冗余数据和噪声,在进行大数据分析任务时不仅会大大增加计算机的内存负荷和处理时间,而且会降低算法的性能[7-8]。特征选择已经被证实在处理高维数据方面是有效且高效的。

近年来,无标签场景下的特征选择引起学者们的广泛关注。无监督特征选择方法根据选择特征的策略可分为过滤方法、包装器方法、混合方法 3 种主要方法。

1)过滤方法根据数据的内在属性评估特征。He 等[9]提出拉普拉斯分数(Laplacian)方法是最经典的过滤方法之一。它根据保留原始数据固有结构的能力,分别计算每个特征的 Laplacian,特征分数越小其局部保持力就越强,越有代表性。Du 等[10]提出的矩阵分解鲁棒无监督特征选择方法(robust unsupervised feature selection via matrix factorization,RUFSM)可以在某范式下同时进行鲁棒判别特征选择和鲁棒聚类,同时保留数据的局部流形结构。Tang 等[11]提出一种通过对偶自表示和流形正则化(dual self-representation and manifold regularization,DSRMR)进行鲁棒无监督特征选择的有效方法。

2)包装器方法使用特定聚类算法的结果评估特征子集。Guo 等[12]提出与嵌入式无监督特征选择(embedded unsupervised feature selection,EUFS)[13]相同的目标函数,其不同之处在于最终模型的损失函数采用 Frobenius 范数代替 l_2 范数,采用 K 均值聚类算法更新参数迭代进行,直到模型收敛。Guo 等[14]提出的依赖引导式无监督特征选择(dependence guided unsupervised feature selection,DGUFS)能够同时执行特征选择和基于 l_2 范数的约束模型聚类。

3)混合方法[15-16]试图综合过滤器和包装器的质量,在效率和有效性之间达成良好的折中。Li 等[17]利用非负谱分析获取更准确的聚类标记指标,同时将聚类标记与特征选择矩阵实现联合迭代学习,给出全新的特征选择算法——非负判别特征选择(nonnegative discriminant feature selection,NDFS)。Yang 等[18]给出的无监督判别的无监督算法(unsupervised discriminative feature selection,UDFS)把判别分析和 $l_{2,1}$ 范数最小化结合形成联合框架,优化计算筛选出在批处理的模式下最具辨别力的特征集合。UDFS、NDFS 方法仅集中于鉴别特征的选择,这是聚类或分类问题中的主要作用。然而,这些方法没有考虑所选择特征的高相关性,对聚类或分类结果具有不利影响。为了选择具有低冗余度的特征,Cai 等[19]提出的多簇特征选择(multi-cluster feature selection,MCFS)算法考虑数据内在的流形结构利用谱分析进行学习,可以更好保留特征的多簇结构。谢娟英等[20]给出基于标准差的谱聚类无监督特征选择方法(feature selection by spectral clustering based on standard deviation,FSSC-SD),通过谱分析方法对各种特征属性进行聚类,然后定义特征重要度评价指标为特征独立性与特征区分度之积,从各聚类类簇中找出重要度最大即最具代表性的特征加入特征子集。

本文通过对类簇中的特征进行等距映射和稀疏学习打分,可以更好地保留数据原始流形结构,促进类簇中价值信息的表达,进而可以选择出该类簇的高代表性的强分类特征,提升准确率。

1 FSFCI 算法

本文针对无标签场景下特征选择方法大多无法很好兼顾到分析结果的准确率和稳定性的问题开展研究,提出新的特征选择算法——基于特征聚类和等距映射的无监督特征选择算法(unsupervised feature selection algorithm based on feature clustering and isometric mapping,FSFCI)。

具体的算法架构见图 1,主要分为等距映射特征聚类簇、学习稀疏系数向量、代表性特征选择 3 个模块。

等距映射特征聚类簇模块首先基于特征的相关性将数据集 $\boldsymbol{X}$ 的特征通过 K 均值聚类聚成 d 个类簇 $\boldsymbol{X}^d$,然后通过等距映射(isometric mapping,ISOMAP)方法使 $\boldsymbol{X}^d$ 形成高维数据的低维嵌入表达 $\boldsymbol{Y}^d$。在 ISOMAP 算法中,首先通过迪杰斯特拉(Dijsktra)算法计算 K 邻接图中样本点之间最短路径来构建距离矩阵 $\boldsymbol{D}$,然后调用多维尺度变换(multiple dimensional scaling,MDS)算法得到类簇的低维表示 $\boldsymbol{Y}^d$。

学习稀疏系数向量模块通过构建 $\boldsymbol{X}^d$ 与 $\boldsymbol{Y}^d$ 的线性回归模型对系数矩阵进行优化,使用具有基数约束的最小角度回归(least angle regression,LARs)算

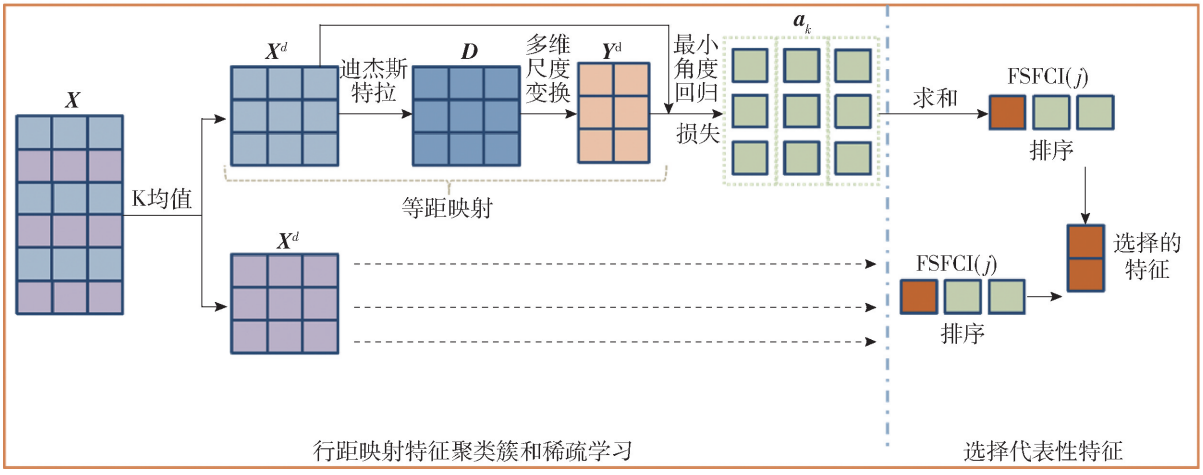

图 1　算法流程

Fig. 1　Algorithm flow

法解决 l_1 正则化回归问题，从而学习到稀疏系数矩阵 $\boldsymbol{a}_k$，$\boldsymbol{a}_k$ 代表了 $\boldsymbol{X}^d$ 与 $\boldsymbol{Y}^d$ 的相关程度表示第 k 个特征的重要性。

代表性特征选择模块将各个特征的稀疏系数根据重要度评分函数 FSFCI(j) 加和排序后，选出该类簇中分数最大的特征作为该类簇的代表性特征加入特征子集中，最终 d 个类簇形成具有 d 个代表特征的特征子集。

1.1　等距映射特征聚类簇

学者们已经提供了许多流形的降维算法。这些方法可以通过对数据的整体构造和几何构造来分解数据并实现降维，由欧氏空间转换到流形空间，最后得到和数据几何构造保持一致的低维嵌入表示。本文的特征选择算法是先利用 ISOMAP 算法来学习对高维数据的低维嵌入，它是一个能够有效掌握高维数据内在几何构造特征的流形学算法。

首先根据特征之间的相关性进行聚类，把具有强依赖性的特征聚成 d 个类簇，每个类簇中有 $d_k(k=1,2,\cdots,d)$ 个特征。本文对各个类簇中样本数据使用 ISOMAP 算法。由于 ISOMAP 算法中采用了 MDS 降维，因此能最大限度地保持数据点的内在几何结构，即保持两点间的测地距离。首先计算出样本点之间的欧氏距离矩阵，构建邻域关系图 $\boldsymbol{G}(\boldsymbol{V},\boldsymbol{E})$，对每个 $\boldsymbol{x}_i(i=1,2,\cdots,N)$ 采用 K 最近邻（K-nearest neighber，KNN）算法计算其 p 近邻 x_{i1}，$x_{i2},\cdots,x_{ip}$，记为 N_j，以点 $\boldsymbol{x}_i$ 为定点，欧氏距离 $d(\boldsymbol{x}_i,\boldsymbol{x}_{ij})$ 为边，建立邻域关系图 $\boldsymbol{G}(\boldsymbol{V},\boldsymbol{E})$。传统的计算测地距离 $\boldsymbol{D}=(d_{ij\_})_{n\times n}$，在 $\boldsymbol{G}(\boldsymbol{V},\boldsymbol{E})$ 中寻找最短路径，公式为

$$d_{ij}=\begin{cases} d_{ij}, & \forall x_j\in N_i \text{ 或 } x_i\in N_j \\ \min\limits_p\{d_{ij},d_{ip}+d_{pj}\}, & \text{其他情况} \end{cases} \tag{1}$$

接着对 $\boldsymbol{D}=(d_{ij})_{n\times n}$ 运用 MDS 方法降维，从而将原始高维的样本数据用更低维的特征向量来加以表达，实现了流形空间与其对应低维空间之间的映射关系，并由此得到对应的高维数据的低维嵌入表达 $\boldsymbol{Y}^d$ 来完成特征选择。

1.2　学习稀疏系数向量

在 1.1 节得到类簇中样本数据的低维嵌入 $\boldsymbol{Y}^d$ 的基础上沿着每个内在维度（$\boldsymbol{Y}^d$ 的每个列）来评估每个特征的重要性。给定一个 $\boldsymbol{y}_k$，通过最小化公式的训练损失来指导获得一个相关的特征子集。

$$\min_{a_k}\|\boldsymbol{y}_k-\boldsymbol{X}^{\mathrm{T}}\boldsymbol{a}_k\|^2+\beta|\boldsymbol{a}_k| \tag{2}$$

式中：$\boldsymbol{a}_k$ 是一个 d_k 维向量，$|\boldsymbol{a}_k|=\sum\limits_{j=1}^{d_k}|a_{k,j}|$ 表示 $\boldsymbol{a}_k$ 的 l_1 范数。由于 l_1 范数的惩罚特性，如果 β 足够大，$\boldsymbol{a}_k$ 的一些系数将缩小到零。$\boldsymbol{a}_k$ 代表不同特征的组合系数矩阵。在最小化过程中，随着 $\boldsymbol{a}_k$ 的变化，$\boldsymbol{X}^{\mathrm{T}}\boldsymbol{a}_k$ 逐渐逼近 $\boldsymbol{y}_k$，从而可以指导算法选择出与 $\boldsymbol{y}_k$ 最相关的特征（对应于 $\boldsymbol{a}_k$ 中的非零系数）。

式(2)中的回归问题，其等效公式为

$$\begin{aligned} &\min_{a_k}\|\boldsymbol{y}_k-\boldsymbol{X}^{\mathrm{T}}\boldsymbol{a}_k\|^2 \\ &\text{s.t.}\quad |\boldsymbol{a}_k|\leqslant\gamma \end{aligned} \tag{3}$$

LARs 算法可用于求解式(3)中的优化问题。LARs 通过指定 $\boldsymbol{a}_k$ 的基数（非零条目数）来控制 $\boldsymbol{a}_k$ 的稀疏性，而不是设置参数 γ，这对于特征选择特别方便。

1.3 代表性特征挑选

考虑从每个类簇中选择一个代表特征来形成特征子集。对于包含 d 个类别的数据集,可以使用 1.2 中讨论的方法来计算 d 个稀疏系数向量。由于类别未知,本算法设 $d=10$。$\boldsymbol{a}_k$ 的基数为 d_k,$\boldsymbol{a}_k$ 的每个条目对应一个特征,每个类簇中选择的所有特征都至少在 $\{\boldsymbol{a}_k\}_{k=1}^{d} \in \mathbb{R}^M$ 中有一个非零系数,实际上,可以使用以下简单而有效的方法从每个类簇中选出代表特征。对于每个类簇中特征 j,本文将该特征的 FSFCI 分数定义为

$$\mathrm{FSFCI}(j)=\sum_{k=1}^{d} a_{k,j} \tag{4}$$

式中 $a_{k,j}$ 是向量 $\boldsymbol{a}_k$ 的第 j 个元素。然后,本文按照 FSFCI 分数降序排列所有特征,并选择每个类簇中评分最大特征作为代表特征,形成特征子集。

1.4 算法思想描述

FSFCI 算法描述如下。

输入:N 个具有 M 特征的数据样本 $\boldsymbol{X}$;最近邻数 p;LARs 基数约束 γ。

输出:选择的 d 个特征。

开始

1) 等距映射特征聚类簇
   ① 根据特征的相关性做 K 均值聚类,使特征聚成 d 个类簇 $\boldsymbol{X}^d$。
   ② 对每个类簇进行 ISOMAP 降维得到数据的低维嵌入表示 $\boldsymbol{Y}^d$。
2) 学习稀疏系数向量
   ① 类簇中建立 $\boldsymbol{X}^d$ 与 $\boldsymbol{Y}^d$ 的线性回归模型,即式(2)。
   ② 使用基数约束为 γ 的 LARs 算法解决式(3)的 l_1 正则回归问题,得到稀疏系数矩阵 $\boldsymbol{a}_k$。
3) 代表性特征挑选
   ① 按照式(4)计算每个类簇中各特征评分并降序排列。
   ② 选出每个类簇中评分最大的特征作为代表特征,形成特征子集。

END

2 实验

2.1 数据集

实验选择了 14 个公开数据集对算法进行测试研究,实验数据集可以从 Feature Selection Datasets, Cancer Gene Expression Data Sets 数据库获取,数据集的详细信息见表 1,其中, warpAR10P、warpPIE10P、YaleB 和 Yale 是人脸数据集,COIL20 是实物数据集,GLIOMA、lymphoma、Prostate-GE 和 Lung-discrete 是生物数据集, RELATHE 和 PCMAC 是文本数据集,Umist 和 USPSdata_20 是手写数字数据集,Isolet 为音频数据集。

表 1 实验数据集描述

Table 1 Description of experimental data sets

数据集	样本个数	特征个数	类别个数
RELATHE	1 427	4 322	2
warpAR10P	130	2 400	10
warpPIE10P	210	2 420	10
GLIOMA	50	4 434	4
COIL20	1 440	1 024	20
YaleB	2 414	1 024	38
Yale	165	1 024	15
lymphoma	96	4 026	9
Isolet	1 560	617	26
Prostate-GE	102	5 966	2
Lung-discrete	73	325	7
Umist	575	644	20
USPSdata_20	1 854	256	10
PCMAC	1 943	3 289	2

2.2 参数设置

为了检验所提出的 FSFCI 算法的性能,实验比较了 FSFCI 算法与其他无监督特征选择算法在选择 2~100 个特征(步长为 2)以及选择所有特征时的实验结果。用作对比的算法包括经典 Laplacian[9]、MCFS[19]、NDFS[17]、UDFS[18]和较新的 FSSC-SD[20]。

本算法实验中均采用欧氏距离计算特征间距离,采用热核相似性度量特征间相似性,近邻数 p 设置为 5,带宽参数 t 设置为 1。NDFS 方法的参数 γ 设置为 108,α 和 β 均设置为 1。UDFS 算法将正则化参数设定为 0.1。

对于所有的实验数据集,特征选择的个数设定为 $\{2,4,6,\cdots,100\}$,采用 KNN 算法对数据集进行了分类,KNN 算法中参数设置 $p=5$。采用十折的交叉检验方式划分训练集与实验集,并采用最大最小化的方式对数据进行标准化。每组实验进行 5 次十折交叉验证,选取 5 次结果的平均值作为实验结果来比较各算法的性能。

2.3 评价标准

评价标准指标包括分类准确率、受试者工作特征曲线下面积、准确性与召回率的调和平均数。分类准确度表示所有样本中完全划分准确的比率。受

试者工作特征曲线下面积表示预测的正例排在负例前面的概率。准确性与召回率之间的调和平均数，只有在准确性与召回率都非常高时，其值才会高。上面的 3 个指标的值越大，说明算法效果越好。

2.4　实验结果分析

本文对比了所提出的 FSFCI 算法与其他 5 种无监督特征选择算法 FSSC-SD[20]、Laplacian[9]、MCFS[19]、NDFS[17] 及 UDFS[18] 特征选择的性能效果，并对性能指标值进行了综合分析。

表 2 为最大分类准确率数据，表 3 为各算法在最大分类准确率时所选特征子集数量数据。其中最优结果用粗体表示，次优用下划线标注。

表 2　比较算法的最大分类准确率

Table 2　Maximum classification accuracy of the compared algorithms　%

数据集	非特征选择	FSSC-SD[20]	Laplacian[9]	MCFS[19]	NDFS[17]	UDFS[18]	FSFCI
RELATHE	77.74	73.86	62.21	60.91	64.81	63.42	**78.94**
warpAR10P	51.54	60.46	28.77	51.69	51.08	43.85	**69.23**
warpPIE10P	94.57	96.48	91.90	96.48	96.48	97.24	**98.57**
GLIOMA	75.00	80.00	63.20	72.20	69.20	68.40	**82.40**
COIL20	99.61	99.44	95.94	99.33	99.00	94.28	**99.83**
YaleB	72.14	66.67	63.09	60.51	65.10	58.96	**72.70**
Yale	70.54	71.11	58.79	64.59	69.43	54.60	**71.99**
lymphoma	92.13	91.62	82.20	92.60	86.13	88.98	**94.69**
Isolet	85.26	84.08	55.54	71.47	81.03	74.65	**86.37**
Prostate-GE	85.13	83.45	83.07	84.56	80.40	84.27	**89.83**
Lung-discrete	86.07	88.39	79.46	86.50	88.82	86.50	**90.36**
Umist	98.33	98.85	96.59	96.91	98.68	94.95	**98.96**
USPSdata-20	93.28	92.99	92.68	89.09	93.24	90.82	**93.44**
PCMAC	70.99	65.45	56.35	60.89	57.11	58.48	**67.14**

表 3　最大分类准确率时所选择的特征数量

Table 3　Number of features selected for maximum classification accuracy

数据集	非特征选择	FSSC-SD[20]	Laplacian[9]	MCFS[19]	NDFS[17]	UDFS[18]	FSFCI
RELATHE	全部特征	97	99	89	97	99	**75**
warpAR10P	全部特征	39	97	97	29	93	**23**
warpPIE10P	全部特征	79	97	**31**	97	85	63
GLIOMA	全部特征	71	63	53	81	57	**45**
COIL20	全部特征	93	99	97	99	**79**	79
YaleB	全部特征	93	97	**63**	99	97	83
Yale	全部特征	61	83	99	95	**43**	97
lymphoma	全部特征	89	93	89	**65**	97	89
Isolet	全部特征	91	99	89	**73**	99	96
Prostate-GE	全部特征	53	69	89	79	85	**5**
Lung-discrete	全部特征	65	95	93	53	77	**49**
Umist	全部特征	**55**	91	99	99	**55**	95
USPSdata-20	全部特征	81	93	95	87	99	**71**
PCMAC	全部特征	81	97	**75**	99	97	95

1）准确率对比

表 2 揭示了无监督特征选择算法不但可以大幅减少特征维度，还可以提高分类准确率。尤其是所提出的 FSFCI 算法，它的最大分类准确率与其他算

法相比较都获得最高水平,且平均最大分类准确率能提高 2.97% ~13.19%,证明了 FSFCI 算法中选取的特征子集更具有效性和代表性。

2)最大分类准确率时选择特征数对比

表 3 显示了数据集通过特征选择,可以选择出远少于原始特征维度的并且具有强分类能力的特征。尤其是本文提出的 FSFCI 算法,在保证最大分类准确率的基础上,平均选择特征数最少为 69,体现了本算法特征选择的良好性能。

3)部分结果可视化展示

图 2 是各算法在 warpAR10P 数据集上进行特征选择后的实验结果。可以看出,FSFCI 算法的各项指标值都显著高于其他比较算法,FSSC-SD[20] 算法次之,紧随其后的是 UDFS[18]、MCFS[19] 和 NDFS[17] 算法,Laplacian[9] 算法在此数据集中表现的效果最差。FSFCI 算法在选择特征数为 23 时,分类准确率达到最高,且 FSFCI 算法选择特征数多于 20 之后,各项指标值均最高。但随着选择特征数的增加,FSFCI 算法指标值有些许下降趋势,分析原因为随着选择特征数的增加,选择了更多的冗余和混杂的特征,导致效果有所下降。

图 3 是各算法在 RELATHE 数据集上进行特征选择后的实验结果。可以看出,FSFCI 算法的各项指标值都显著高于其他比较算法,FSSC-SD[20] 算法次之,紧随其后的是 UDFS[18] 和 NDFS[17] 算法,MCFS[19] 和 Laplacian[9] 算法在此数据集中表现的效果较差。FSFCI 算法在选择特征数为 75 时,分类准确率达到最高,且 FSFCI 算法在选择特征数多于 20 后,各项指标值均超过其他算法。随着选择特征数的增加,该算法的各指标值均平稳增加,体现了本算法的良好性能。

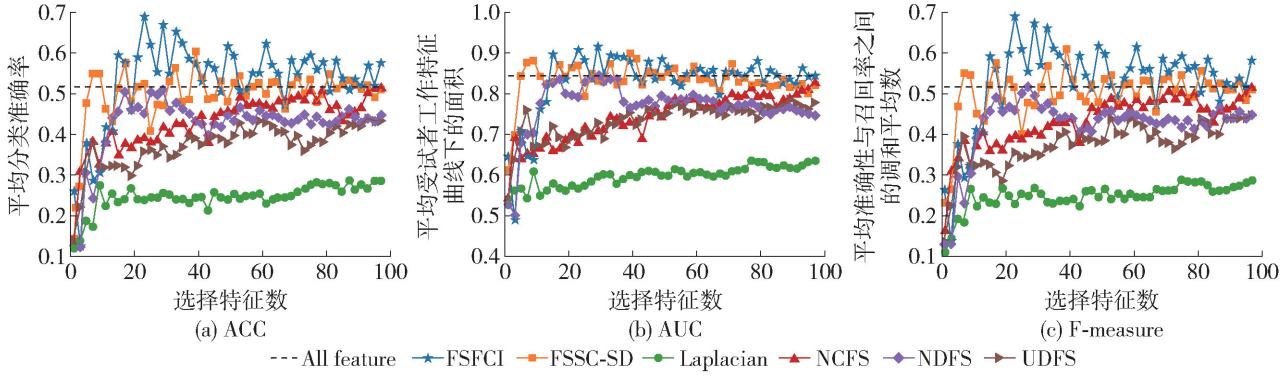

图 2 warpAR10P 数据集特征选择结果对比

Fig. 2 Comparison of warpAR10P data set feature selection results

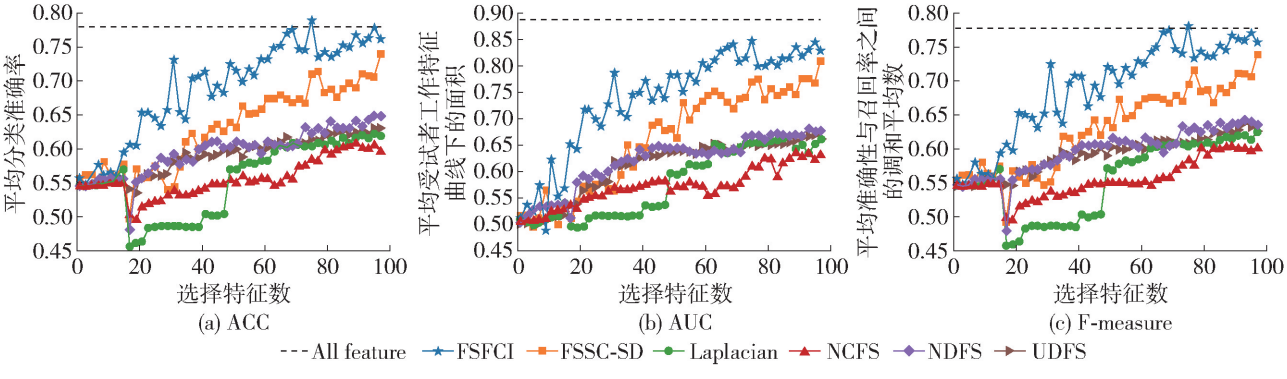

图 3 RELATHE 数据集特征选择结果对比

Fig. 3 Comparison of RELATHE data set feature selection results

图 4 是各算法在 GLIOMA 数据集上进行特征选择后的实验结果图。可以看出,FSFCI 算法的各项指标值都显著高于其他比较算法,FSSC-SD[20] 和 MCFS[19] 算法次之,紧随其后的是 UDFS[18] 和 NDFS[17] 算法,Laplacian[9] 算法在此数据集中表现的效果最差。FSFCI 算法当选择特征数为 45 时,分类精确度获得最高。随着选择特征数的增加,FSFCI 算法指标值也有些许下降的趋势,分析原因为随着选择特征数的增加,选择了更多的冗余和混杂的特征,导致效果有所下降。

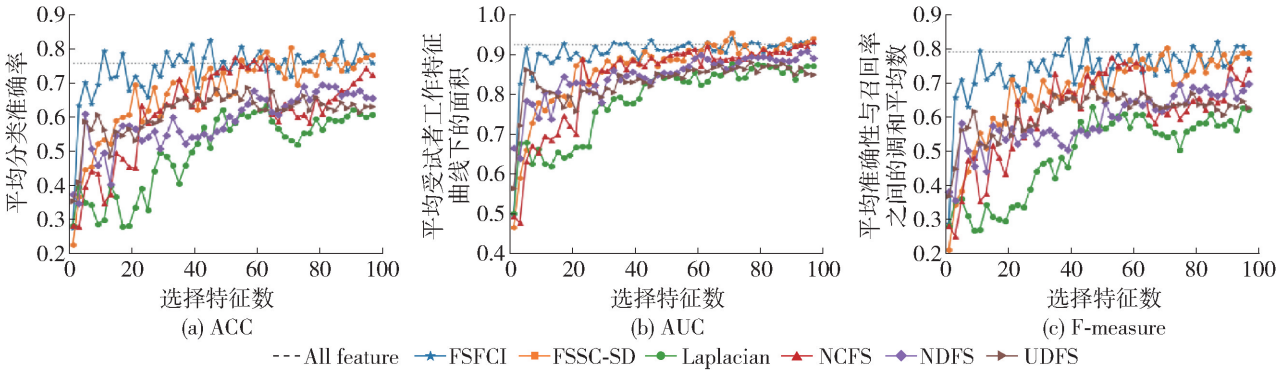

图 4　GLIOMA 数据集特征选择结果对比

Fig. 4　Comparison of GLIOMA data set feature selection results

图 5 是各算法在 COIL20 数据集上进行特征选择后的实验结果图。可以看出，FSFCI 算法的各项指标值都显著高于其他比较算法，FSSC-SD[20]、NDFS[17] 和 UDFS[18] 算法次之，MCFS[19]、Laplacian[9] 算法在此数据集中表现的效果较差。FSFCI 算法在选择特征数为 79 时，分类准确率达到最高，且 FSFCI 算法在选择特征数多于 15 后，各项指标值均已经超过其他特征选择算法的结果。

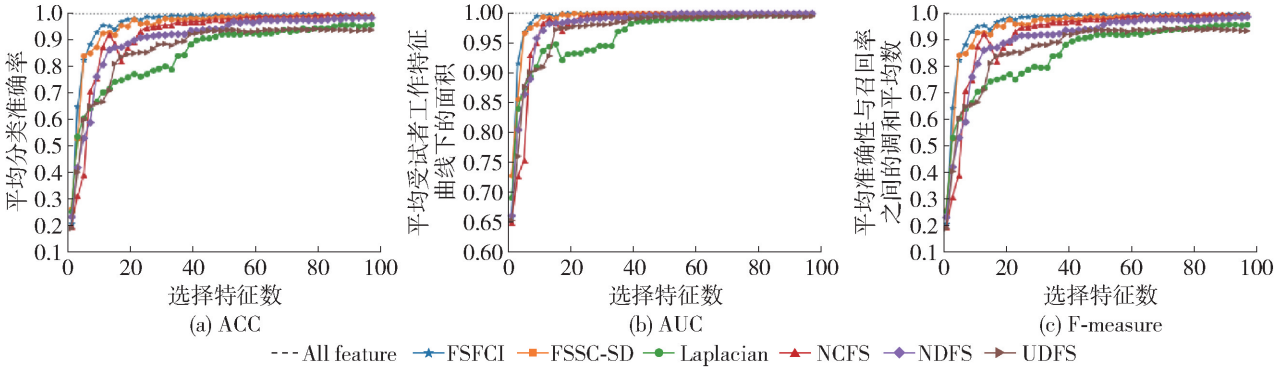

图 5　COIL20 数据集特征选择结果对比

Fig. 5　Comparison of COIL20 data set feature selection results

综合实验结果来看，FSFCI 算法能够选择出具有强分类能力的特征子集，有效地提升实验的分类准确率，并且有很好的泛化能力。

3　结论

1）本文针对无标签场景下特征选择方法大多无法很好兼顾到选择结果的准确率和稳定性的问题，提出了面向高维多源数据集的 FSFCI 算法，算法根据等距映射定义新的特征得分函数，为各特征簇中特征进行打分，挑选类簇中得分最高的特征作为代表特征。实验结果表明，FSFCI 算法可以取得比经典流行的 FSSC-SD[20]、Laplacian[9]、MCFS[19]、NDFS[17]、UDFS[18] 更好的效果，尤其是准确率提升明显，说明它对高维数据具有较好的处理能力。同时在多个高维数据集上的良好表现，表明了 FSFCI 算法对数据集的依赖性较小，因而具有一定的普适性。

2）本文提出的 FSFCI 算法在多种高维特征数据集上均有很好的效果，但是仍存在可以改进的空间。一方面可以从提高算法的最大分类准确率和减小算法的时间成本 2 个方向进一步优化算法。另一方面，由于现有的大多数基于聚类的特征选择技术都需要额外指定选择特征的数量以及大量的参数，而实际中，无法确定每个数据集所对应的最优选择特征数量与最适参数，因此可以在自适应的方向做进一步研究。

参考文献：

[1] SONG J, GUO Y, GAO L, et al. From deterministic to generative: multimodal stochastic RNNs for video captioning[J]. IEEE transactions on neural networks and learning systems, 2018, 30(10): 3047-3058.

[2] LI X, CHEN M, NIE F, et al. A multiview-based parameter free framework for group detection[C]// Thirty-

first AAAI Conference on Artificial Intelligence. Palo Alto, California, USA: Association for the Advancement of Artificial Intelligence, 2017: 4147-4153.

[3] LI X, CHEN M, NIE F, et al. Locality adaptive discriminant analysis[C]// Proceedings of the Twenty-sixth International Joint Conference on Artificial Intelligence. Freiburg, Germany: IJCAI, 2017: 2201-2207.

[4] ZHANG R, TONG H. Robust principal component analysis with adaptive neighbors[J]. Advances in neural information processing systems, 2019, 32(8): 6959-6967.

[5] WANG F, WANG Q, NIE F, et al. Unsupervised linear discriminant analysis for jointly clustering and subspace learning[J]. IEEE Transactions on Knowledge and Data Engineering, 2019, 33(3): 1276-1290.

[6] SONG J, GAO L, NIE F, et al. Optimized graph learning using partial tags and multiple features for image and video annotation[J]. IEEE Transactions on Image Processing, 2016, 25(11): 4999-5011.

[7] LEE P Y, LOH W P, CHIN J F. Feature selection in multimedia: the state-of-the-art review[J]. Image and Vision Computing, 2017, 67: 29-42.

[8] ZHANG R, NIE F, LI X, et al. Feature selection with multi-view data: a survey[J]. Information Fusion, 2019, 50: 158-167.

[9] HE X, CAI D, NIYOGI P. Laplacian score for feature selection[J]. Advances in Neural Information Processing Systems, 2005, 18(5): 504-514.

[10] DU S, MA Y, LI S, et al. Robust unsupervised feature selection via matrix factorization[J]. Neurocomputing, 2017, 241: 115-127.

[11] TANG C, LIU X, LI M, et al. Robust unsupervised feature selection via dual self-representation and manifold regularization[J]. Knowledge-Based Systems, 2018, 145: 109-120.

[12] GUO J, GUO Y, KONG X, et al. Unsupervised feature selection with ordinal locality[C]// 2017 IEEE International Conference on Multimedia and Expo (ICME). Piscataway, NJ: IEEE, 2017: 1213-1218.

[13] WANG S, TANG J, LIU H. Embedded unsupervised feature selection[C]// Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto, California: Association for the Advancement of Artificial Intelligence, 2015: 470-476.

[14] GUO J, ZHU W. Dependence guided unsupervised feature selection[C]// Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto, California: Association for the Advancement of Artificial Intelligence, 2018: 2232-2239.

[15] WANG X, ZHANG X, ZENG Z, et al. Unsupervised spectral feature selection with l1-norm graph[J]. Neurocomputing, 2016, 200: 47-54.

[16] PENG C, KANG Z, YANG M, et al. Feature selection embedded subspace clustering[J]. IEEE Signal Processing Letters, 2016, 23(7): 1018-1022.

[17] LI Z, YANG Y, LIU J, et al. Unsupervised feature selection using nonnegative spectral analysis[C]// Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto, California: Association for the Advancement of Artificial Intelligence 2012: 1026-1032.

[18] YANG Y, SHEN H T, MA Z, et al. L2, 1-norm regularized discriminative feature selection for unsupervised[C]// Proceedings of the Twenty-second International Joint Conference on Artificial Intelligence. Freiburg, Germany: IJCAI, 2011: 1589-1594.

[19] CAI D, ZHANG C, HE X. Unsupervised feature selection for multi-cluster data[C]// Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: Association for Computing Machinery, 2010: 333-342.

[20] 谢娟英, 丁丽娟, 王明钊. 基于谱聚类的无监督特征选择算法[J]. 软件学报, 2020, 31(4): 1009-1024.
XIE J Y, DING L J, WANG M Z. Unsupervised feature selection algorithm based on spectral clustering[J] Journal of Software, 2020, 31(4): 1009-1024. (in Chinese)

(责任编辑 郑筱梅)