软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2020,31(4):1009−1024 [doi: 10.13328/j.cnki.jos.005927]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于谱聚类的无监督特征选择算法*

谢娟英[1], 丁丽娟[1,3], 王明钊[2]

[1](陕西师范大学 计算机科学学院,陕西 西安 710062)

[2](陕西师范大学 生命科学学院,陕西 西安 710062)

[3](武警工程大学 信息工程学院,陕西 西安 710086)

通讯作者: 谢娟英, E-mail: xiejuany@snnu.edu.cn

摘 要: 基因表达数据具有高维小样本特点,包含了大量与疾病无关的基因,对该类数据进行分析的首要步骤是特征选择.常见的特征选择方法需要有类标的数据,但样本类标获取往往比较困难.针对基因表达数据的特征选择问题,提出基于谱聚类的无监督特征选择思想 FSSC(feature selection by spectral clustering).FSSC 对所有特征进行谱聚类,将相似性较高的特征聚成一类,定义特征的区分度与特征独立性,以二者之积度量特征重要性,从各特征簇选取代表性特征,构造特征子集.根据使用的不同谱聚类算法,得到 FSSC-SD(FSSC based on standard deviation)、FSSC-MD(FSSC based on mean distance)和 FSSC-ST(FSSC based on self-tuning)这 3 种无监督特征选择算法.以 SVMs (support vector machines)和 KNN(*K*-nearest neighbours)为分类器,在 10 个基因表达数据集上进行实验测试.结果表明,FSSC-SD、FSSC-MD 和 FSSC-ST 算法均能选择到具有强分类能力的特征子集.

关键词: 谱聚类;无监督特征选择;特征独立性;特征区分度;特征重要度

中图法分类号: TP181

中文引用格式: 谢娟英,丁丽娟,王明钊.基于谱聚类的无监督特征选择算法.软件学报,2020,31(4):1009−1024. http://www.jos.org.cn/1000-9825/5927.htm

英文引用格式: Xie JY, Ding LJ, Wang MZ. Spectral clustering based unsupervised feature selection algorithms. Ruan Jian Xue Bao/Journal of Software, 2020,31(4):1009−1024 (in Chinese). http://www.jos.org.cn/1000-9825/5927.htm

Spectral Clustering Based Unsupervised Feature Selection Algorithms

XIE Juan-Ying[1], DING Li-Juan[1,3], WANG Ming-Zhao[2]

[1](School of Computer Science, Shaanxi Normal University, Xi'an 710062, China)

[2](College of life Sciences, Shaanxi Normal University, Xi'an 710062, China)

[3](College of Information Engineering, Engineering University of PAP, Xi'an 710086, China)

Abstract: Gene expression data usually comprise small number of samples with tens of thousands of genes. There are a large number of genes unrelated to diseases in this kind of data. The primary task is to detect those key essential genes when analyzing this kind of data.

* 基金项目: 国家自然科学基金(61673251); 陕西省科技攻关重点项目(2018ZDXMSF-079); 国家重点研发计划(2016YFC0901900); 科技成果转化培育项目(GK201806013); 中央高校基本科研业务费专项资金(GK201701006); 研究生培养创新基金(2015CXS028, 2016CSY009, 2018TS078)

Foundation item: National Natural Science Foundation of China (61673251); Key Projects of Science and Technology Research in Shaanxi Province (2018ZDXMSF-079); National Key Research and Development Program of China (2016YFC0901900); Scientific and Technological Achievements Transformation and Cultivation Funds of Shaanxi Normal University (GK201806013); Fundamental Research Funds for the Central Universities (GK201701006); Innovation Funds of Graduate Programs at Shaanxi Normal University (2015CXS028, 2016CSY009, 2018TS078)

本文由"非经典条件下的机器学习方法"专题特约编辑高新波教授、黎铭教授、李天瑞教授推荐.

收稿时间: 2019-05-31; 修改时间: 2019-07-29; 采用时间: 2019-09-20; jos 在线出版时间: 2020-01-10

CNKI 网络优先出版: 2020-01-14 09:53:31, http://kns.cnki.net/kcms/detail/11.2560.TP.20200114.0953.012.html

The common feature selection algorithms depend on labels of data, but it is very difficult to get labels for data. To overcome the challenges, especially for gene expression data, the unsupervised feature selection idea is proposed, named as FSSC (feature selection by spectral clustering). FSSC groups all of features into clusters by a spectral clustering algorithm, so that similar features are in same clusters. The feature discernibility and independence are defined, and the feature importance is defined as the product of its discernibility and independence. The representative feature is selected from each cluster to construct the feature subset. According to the spectral clustering algorithms used in FSSC, three kinds of unsupervised feature selection algorithms named as FSSC-SD (FSSC based on standard deviation), FSSC-MD (FSSC based on mean distance) and FSSC-ST (FSSC based on self-tuning) are developed. The SVM (support vector machines) and KNN (*K*-nearest neighbors) classifiers are adopted to test the performance of the selected feature subsets in experiments. Experimental results on 10 gene expression datasets show that FSSC-SD, FSSC-MD, and FSSC-ST algorithms can select powerful features to classify samples.

Key words: spectral clustering; unsupervised feature selection; feature independence; feature discernibility; feature importance

生物测序技术的迅速发展实现了大规模基因表达数据的自动获取,为癌症等疾病的发病机理和诊断研究提供了新途径[1−3].然而,基因表达数据具有高维小样本特点,包含了大量与疾病无关的基因(冗余基因)[4−6].因此,选取具有高分类信息的基因子集是分析基因表达数据的首要任务[5,7].特征选择可以筛选出与分类任务高度相关的基因,提高分类准确率[8].

特征选择是从原始特征集合中选取具有强分类信息且尽可能相互独立的特征构成特征子集,以尽可能地保留原始系统的分类信息且包含尽可能少的特征,从而达到去除冗余特征、提高分类准确率的目的.根据与分类器的关系,特征选择算法分为 Filter、Wrapper 和 Embedded 方法[9−11].根据是否使用类标信息,特征选择算法又分为有监督特征选择方法和无监督特征选择方法.

1 相关工作介绍

有监督特征选择方法通过计算特征与类标的相关性进行特征选择,如 Relief 算法[12]、mRMR(minimal redundancy-maximal relevance)算法[13]、CFS(correlation-based feature selection for machine learning)算法[14]等.然而,样本类标往往很难获得,因此无监督特征选择算法引起了研究者的关注.

Dash 等人[15]提出基于熵排序的无监督特征选择算法,利用信息熵度量特征重要性程度,从而选择最优特征子集;徐峻岭等人[16]提出基于互信息的无监督特征选择算法,利用互信息定义特征的相关度与冗余度,综合考虑特征的相关度与冗余度来评价特征重要性;张莉等人[17]提出基于 *K*-均值聚类的无监督特征选择算法,利用特征对聚类结果的影响以及特征之间的相关性作为特征选择的判别标准;He 等人[18]针对无监督特征选择算法多是 Wrapper 方法,提出独立于任何学习算法的 Filter 特征选择方法——Laplacian score 无监督特征选择算法,利用同类样本距离更近原理,对每个特征计算其拉普拉斯分数以反映其局部保持能力,Laplacian score 越小的特征其局部保持能力越强,重要度越高,越具有代表性,需要说明的是,该算法也可以通过监督的方式执行;Cai 等人[19]提出多类簇无监督特征选择算法 MCFS(multi-cluster feature selection),使用谱聚类技术,然后求解带有 L1 正则项的最小二乘问题,并定义特征的 MCFS score,选择 MCFS score 位于最前面的若干个特征,使选择的特征既能保留更多数据类簇结构,又能覆盖所有可能类簇的特征;王连喜等人[20]提出了一种基于聚类集成的特征选择算法,利用聚类算法将冗余特征聚成一类簇,然后从各类簇挑选代表性特征构成最优特征子集;Zhao 等人[21]基于谱图理论提出 SPEC 算法,以特征值的分布与目标的概念是否一致作为评价准则进行特征选择;我们团队提出基于密度峰值的无监督特征选择算法,分别定义了特征密度与特征距离,以二者之积度量特征的重要性[8,22];He 等人[23]提出基于决策图的无监督特征选择算法 DGFS(decision graph based feature selection),定义特征的局部密度、判别距离和决策图得分,利用局部密度度量特征代表性,利用判别距离度量特征之间的冗余性与相似性,以决策图得分作为评价标准进行特征选择,决策图得分较高的若干特征构成特征子集;我们团队提出基于基因密度峰值发现的结肠癌患者诊断基因标志物识别算法[24],定义基因局部密度和距离,以密度峰值点基因作为结肠癌患者的识别基因;鲁棒的无监督特征选择方法 RUFS(robust unsupervised feature selection)[25]不同于传统无监督特征选择方法,通过局部学习正则化的鲁棒非负矩阵分解,学习样本的伪类簇标签,在标签学习过程中,通过

鲁棒加入 $l_{2,1}$ 范数最小化,同时完成特征选择.RUFS 算法在标签学习和特征学习过程中引入了 $l_{2,1}$ 范数,能够有效地处理异常点和噪音,并能有效去除冗余和噪音特征,兼具鲁棒非负矩阵分解、局部学习和鲁棒特征学习的优势.同时,RUFS 算法基于有限内存投影,采用迭代算法解决了算法伸缩性问题;非负判别特征选择算法 NDFS (nonnegative discriminant feature selection)[26]采用谱聚类学习样本类标,在学习样本类标过程中完成特征选择.类簇标签和特征选择矩阵的联合学习使 NDFS 算法能够选择最具鉴别性的特征.算法中为了学到更准确的类别标签,对类指示器添加了非负约束,为了减少冗余甚至噪声特征,在目标函数中加入 $l_{2,1}$ 范数最小化约束,保证特征选择矩阵的行稀疏性.算法利用判别信息和特征关联来选择更好的特征子集,并设计了一种简单、有效的迭代算法来优化目标函数.因为谱聚类的强大优势,基于谱聚类思想的特征选择方法得到越来越多学者的关注[17,18,20,21,27,28].

本文借助谱聚类算法能够发现任意形状类簇,收敛于全局最优解的性能,提出基于谱聚类的无监督特征选择思想 FSSC(feature selection by spectral clustering).首先对特征进行谱聚类,使相似(具有强冗余性)的特征聚在同一类簇,定义特征区分度和独立性,以特征区分度与独立性之积量化其重要性,选择各类簇最重要的特征代表该类簇特征,各类簇的代表特征构成特征子集.利用 SC_SD[29](spectral clustering based on standard deviation)、SC_MD[29](spectral clustering based on mean distance)、self-tuning[30]算法进行特征谱聚类,得到 3 种不同谱特征选择算法:FSSC-SD(FSSC based on SC_SD)、FSSC-MD(FSSC based on SC_MD)和 FSSC-ST(FSSC based on self-tuning).与其他无监督特征选择算法相比,所提出的算法同时考虑了特征区分度和独立性,能够选择到代表性强的特征子集.10 个癌症基因数据集的实验测试结果表明,FSSC-SD、FSSC-MD 和 FSSC-ST 算法均能选择到具有丰富分类信息的关键基因,为癌症发病机理、早期诊断、治疗等提供支撑与基础.

2 FSSC 无监督特征选择

特征选择旨在选择具有高分类信息且相互之间低冗余的特征构成特征子集.本文利用谱聚类算法能够发现任意形状类簇,收敛于全局最优解的优势,提出谱聚类特征选择思想 FSSC,以期选择既具有强区分能力,又彼此之间相互独立的特征构成特征子集.FSSC 通过对所有特征进行谱聚类,将相似(冗余)特征聚到同一类簇,定义特征标准差为特征区分度,定义特征与簇内其他区分度更好特征的 Pearson 相关系数和的倒数为特征独立性,定义特征重要性为其区分度与独立性之积,选择各类簇最重要的特征代表该类簇,所有代表特征构成特征子集.其思想框架如图 1 所示.

Fig.1 The frame of proposed FSSC algorithm

图 1 提出的 FSSC 算法框架

2.1 特征谱聚类

聚类根据某种相似性原则将数据对象划分为不同类簇,簇内对象相似性较高,但与其他簇对象相似性较低.传统的 *K*-means 等聚类算法适合发现球状簇,无法发现非凸状的簇[31,32].谱聚类算法以谱图理论为基础,将样本聚类问题转化为以样本为顶点、样本间相似性为顶点连接边权重的带权无向图的划分问题.谱聚类算法能够发现任意形状的簇,且收敛于全局最优解[33].因此,对特征进行谱聚类,有助于揭示特征之间的内在联系,发现真正的特征簇.

对特征进行谱聚类,即以特征为顶点、特征间相似性为顶点连接边权重,将特征聚类问题转换为特征图划分问题,分别采用 SC_SD[29](spectral clustering based on standard deviation)、SC_MD[29](spectral clustering based on mean distance)、self-tuning[30]算法对特征进行谱聚类,得到 FSSC-SD(FSSC based on SC_SD)、FSSC-MD (FSSC based on SC_MD)和 FSSC-ST(FSSC based on self-tuning)这 3 种谱特征选择算法.其中,self-tuning 算法[30]

是一种自适应的谱聚类算法,其对传统谱聚类算法计算亲和矩阵的全局尺度参数σ不能准确体现数据集真实分布信息的缺陷进行了改进,提出了样本 i 的局部尺度参数σ_i,定义σ_i为样本 i 到其第 p 个近邻的欧氏距离,采用样本 i,j 的局部尺度参数σ_i,σ_j计算其亲和系数 $A_{i,j}$.SC_SD 和 SC_MD 谱聚类算法是对 self-tuning 谱聚类算法的改进[29],针对 self-tuning 谱聚类算法的局部尺度参数会受离群点影响的问题,提出的两种完全自适应的谱聚类算法.SC_SD 依据样本 i 标准差 $std_i\left(=\sqrt{\frac{1}{N-1}\sum_{j=1}^{N}d_{i,j}^2}\right)$,$d_{i,j}$是样本 i,j 的欧氏距离,N 为数据集样本数,定义其完全自适应的局部尺度参数 $\sigma_{SD\_i}=\sqrt{\frac{1}{S_i-1}\sum_{j=1}^{S_i}d_{j,i}^2}$, S_i为样本 i 对应邻域半径为 std_i 的邻域内样本数.SC_MD 依据样本 i 与数据集其余样本欧氏距离的均值 $d_{mean\_i}\left(=\frac{1}{N-1}\sum_{j=1,j\neq i}^{N}d_{i,j}\right)$ 来定义样本 i 的局部尺度参数 $\sigma_{MD\_i}=\sqrt{\frac{1}{M_i-1}\sum_{j=1}^{M_i}d_{j,i}^2}$, M_i为样本 i 对应邻域半径为 d_{mean_i}的邻域内样本数.

2.2 特征重要度

给定训练数据集 $\boldsymbol{D}\in\Re^{n\times d}$, 其中,$n$ 和 d 分别表示样本数和特征维数.用 $\boldsymbol{f}_1,\boldsymbol{f}_2,...,\boldsymbol{f}_i,...,\boldsymbol{f}_d$ 表示 d 个特征向量,则 $\boldsymbol{D}=[\boldsymbol{f}_1,\boldsymbol{f}_2,...,\boldsymbol{f}_i,...,\boldsymbol{f}_d]$且 $\boldsymbol{f}_i\in\Re^n$;用 $\boldsymbol{x}_1,\boldsymbol{x}_2,...,\boldsymbol{x}_j,...,\boldsymbol{x}_n$ 表示 n 个样本,则 $\boldsymbol{x}_j\in\Re^d$ 且 $\boldsymbol{D}=[\boldsymbol{x}_1;\boldsymbol{x}_2;...;\boldsymbol{x}_j;...;\boldsymbol{x}_n]$.

定义 1(特征区分度(feature discernibility)). 一个区分能力强的特征对不同类样本的取值往往差异很大,因此具有较大方差(或标准差),本文用特征标准差度量特征的类别区分能力.故定义特征 f_i 的区分度 dis_i 为其标准差 std_i,如式(1)所示.

$$dis_i=std_i=\sqrt{\frac{1}{n-1}\sum_{j=1}^{n}\left(f_{ji}-\frac{1}{n}\sum_{j=1}^{n}f_{ji}\right)^2},\quad i=1,2,...,d;\ j=1,2,...,n \tag{1}$$

f_{ji} 表示样本 j 在第 i 个特征的取值,std_i 表示第 i 个特征的标准差(standard deviation).

定义 2(特征独立性(feature independence)). Pearson 相关系数可以度量两变量之间的相关性,两变量的 Pearson 相关系数绝对值越小,则其越不相关.特征选择的目的是选择区分能力强,且彼此不相关的特征构成特征子集,剔除不相关和冗余特征.以特征子集中的特征来表达样本,不仅可以保持和提高系统分类能力,且能使原系统得到简化.因此,本文以 Pearson 相关系数度量特征独立性,定义特征与同类簇区分能力更强特征的 Pearson 相关系数绝对值和的倒数为特征独立性.对区分度最大的特征,定义其独立性为与本簇最不相关特征的 Pearson 相关性绝对值的倒数.特征独立性定义见式(2),其中,M_j 是特征 f_i 所在的特征类簇.特征 f_i,f_k 间的 Pearson 相关系数定义见式(3).

$$ind_i=\begin{cases}\dfrac{1}{\min\limits_{k\in M_j}\left|r_{f_i,f_k}\right|}, & dis_i=\max\limits_{j=1,...,d}\{dis_j\}\\ \dfrac{1}{\sum\limits_{k:dis_k>dis_i,k\in M_j}\left|r_{f_i,f_k}\right|}, & \text{otherwise}\end{cases} \tag{2}$$

$$r_{f_i,f_k}=\frac{\sum_{j=1}^{n}(f_{ji}-\overline{f_i})(f_{jk}-\overline{f_k})}{\sqrt{\sum_{j=1}^{n}(f_{ji}-\overline{f_i})^2(f_{jk}-\overline{f_k})^2}},\quad i,k=1,2,...,d;\quad j=1,2,...,n \tag{3}$$

式(2)所示特征独立性定义保障了:若第 i 特征与区分能力比它强的特征越不相关,则其独立性越强.区分度最大特征的独立性定义保障了区分能力最强的特征的独立性最强,这样保障了数据集中区分能力最强的特征一定会被选择到特征子集.

定义 3(特征重要度(feature importance)). 特征 f_i 的重要度 $score_i$ 定义为特征区分度与特征独立性之积,见式(4),$score_i$ 越大,特征 f_i 越重要.

$$score_i=dis_i \times ind_i \tag{4}$$

2.3 算法思想描述

输入:训练数据集 $\boldsymbol{D} \in \Re^{n\times d}$, n 为训练样本数,d 为特征数;被选特征子集规模 k;

输出:特征子集 S.

BEGIN

a) 初始化被选特征子集 $S=\varnothing$,全部特征集合为 F;

b) 对全部特征分别采用 SC_SD、SC_MD、self-tuning 算法进行谱聚类,得到 k 个特征簇;

c) 利用公式(4)计算各特征的 *score* 值,从各特征簇选取 *score* 值最大的特征加入特征子集 S;

d) 输出特征子集 S.

END

2.4 算法时间复杂度分析

本文算法的时间消耗主要在步骤 b)的特征谱聚类和步骤 c)的基于特征重要度的特征选择.假设训练数据集包含 n 个样本,每个样本的维数为 d.步骤 b)的特征谱聚类的时间复杂度为 $O(d^2)$.在步骤 c)的基于特征重要度的特征选择过程中,计算特征辨识度的时间复杂度是 $O(n^2d)$,计算特征独立性的时间复杂度是 $O(nd^2)$,计算特征 *score* 值并对其降序排序的时间复杂度是 $O(d\log d)$,由于 $n<<d$,故步骤 c)的时间复杂度为 $O(d^2)$.因此,本文基于谱聚类算法的无监督特征选择算法的时间复杂度为 $O(d^2)$.

3 实验结果与分析

实验采用 10 个常用基因数据集对算法进行测试,实验使用数据集可从 Broad Institute Genome Data Analysis Center(http://portals.broadinstitute.org/cgi-bin/cancer/datasets.cgi)和 Gene Expression Model Selector (http://www.gems-system.org/)获取.数据集详细描述见表 1.

Table 1 The descriptions of datasets used in experiments

表 1 实验数据集描述

数据集	特征数	样本数	类簇数
Colon	2 000	62	2
SRBCT	2 308	83	4
Lymphoma	4 026	45	2
Leukemia	7 129	72	2
DLBCL Tumor	7 129	77	2
Carcinoma	7 457	36	2
CNS	7 129	90	2
LungCancer-Michigan	7 129	96	2
Leukemia_MLL	12 582	72	3
ALL1	12 625	128	2

3.1 实验设计

为了验证所提出的 3 种谱聚类特征选择算法 FSSC-SD、FSSC-MD 和 FSSC-ST 的性能,实验比较了这 3 种算法与基于决策图的无监督特征选择算法 DGFS(decision graph-based feature selection)[23]、多类簇无监督特征选择算法 MCFS(multi-cluster feature selection)[19]、Laplacian 分值特征选择算法(Laplacian score for feature selection)[18]、鲁棒的无监督特征选择方法 RUFS(robust unsupervised feature selection)[25]以及非负判别特征选择算法 NDFS(nonnegative discriminant feature selection)[26]在表 1 数据集的实验结果.其中,在 FSSC-ST 算法中,高斯核函数参数设置为经验值 7,对比算法 DGFS 采用欧式距离计算特征间距离并升序排序,截断距离 d_c 设置为特征总数 2%位置处的距离值;对比算法 Laplacian、RUFS 和 NDFS 采用余弦相似性度量特征相似性,且近邻

数 K 均设置为 5,NDFS 算法的正则化参数设置为 0.1.

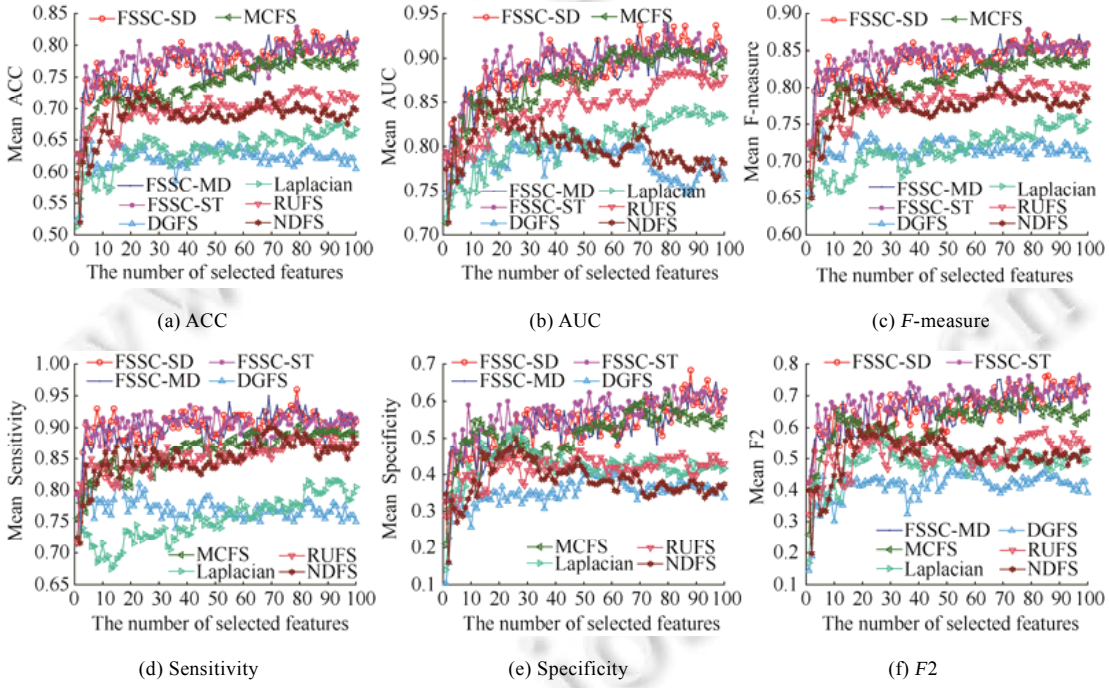
实验采用 10 折交叉验证方法划分训练集与测试集,缺失数据采用类内均值填充,为避免特征间不同量纲对实验结果的影响,采用最大最小化方法标准化数据,采用 SVM 和 KNN(K=1)两种分类器,SVM 分类器采用林智仁等人开发的 SVM 工具箱 Libsvm[34].其中,核函数采用线性核函数,惩罚因子 C 取 20,其余参数均取默认值.以 5 次 10 折交叉验证实验结果的平均值比较各算法的性能,评价准则采用分类正确率 ACC、AUC(或 MAUC)、$F2$[35]、F-measure、Sensitivity 和 Specificity.其中,$F2$ 是针对不平衡数据的评价方法,可以避免 ACC 不适合不平衡数据与 F-measure 主要强调分类器对正类样本识别能力的缺陷[35].MAUC 是 AUC 对多类问题的推广.实验代码使用 MATLAB R2017b 实现;实验环境为 Win10 64bit 操作系统,8GB 内存,Intel(R) Core(TM) i5-6600 CPU @3.30GHz 3.31GHz.

3.2 实验结果比较

本节比较提出的 FSSC-SD、FSSC-MD、FSSC-ST 算法与无监督特征选择算法 DGFS、MCFS、Laplacian、RUFS 及 NDFS 在表 1 所示数据集选择的基因子集的性能,比较各算法选择的基因子集对应分类器的各指标值.

3.2.1 平均实验结果

以 Colon、Carcinoma、ALL1 和 DLBCL-Tumor 数据集为例,对比各算法采用 KNN 的实验结果.图 2~图 5 分别是各算法在 Colon、ALL1、Carcinoma 和 DLBCL-Tumor 数据集对应不同特征子集的实验结果.



(a) ACC (b) AUC (c) F-measure

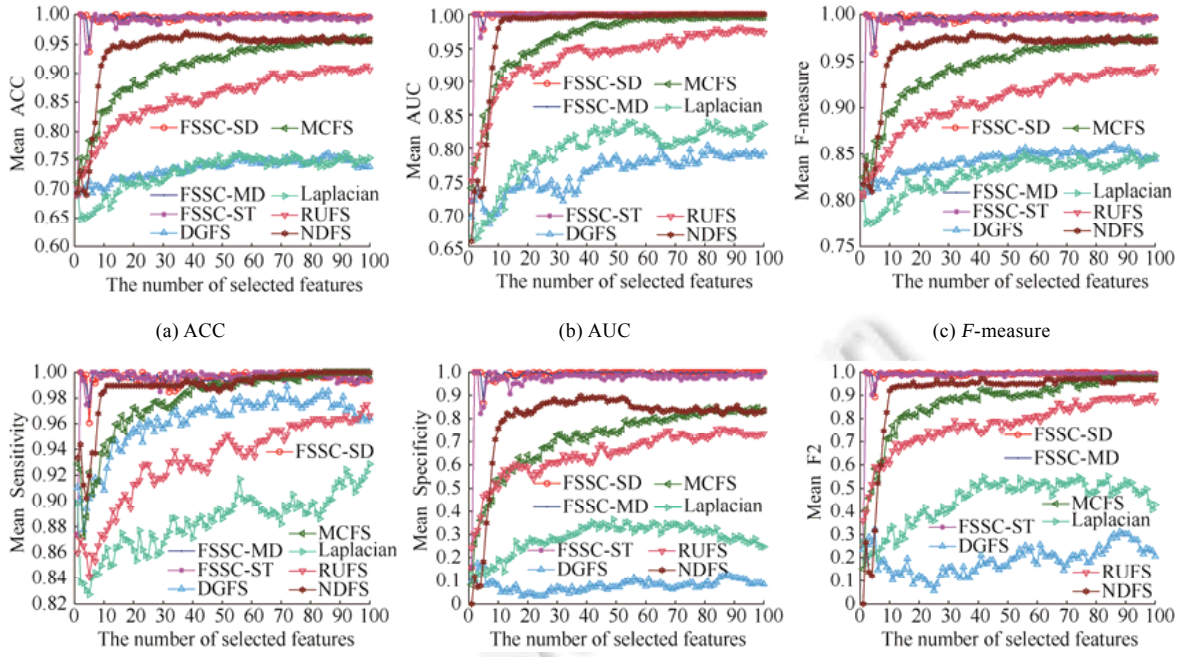
(d) Sensitivity (e) Specificity (f) $F2$

Fig.2 The average indexes of KNN classifier of each algorithm for different feature subsets on Colon dataset

图 2 各算法在 Colon 数据集对应不同特征子集的 KNN 分类器平均指标值

图 2 所示实验结果显示,本文算法 FSSC-SD、FSSC-ST 和 FSSC-MD 选择的特征子集的 KNN 分类器的各指标值绝对地优于对比算法.MCFS、RUFS 算法次之,接着是 NDFS 算法,DGFS 算法与 Laplacian 算法选择的特征子集的 KNN 分类器的性能最差.

图 3 所示实验结果显示,本文所提出的 FSSC-SD、FSSC-MD、FSSC-ST 算法选择的特征子集的分类性能最优,接着是 NDFS 算法,MCFS 和 RUFS 算法选择的特征子集的分类性能居中,DGFS 算法与 Laplacian 算法选择的特征子集的分类能力最差.

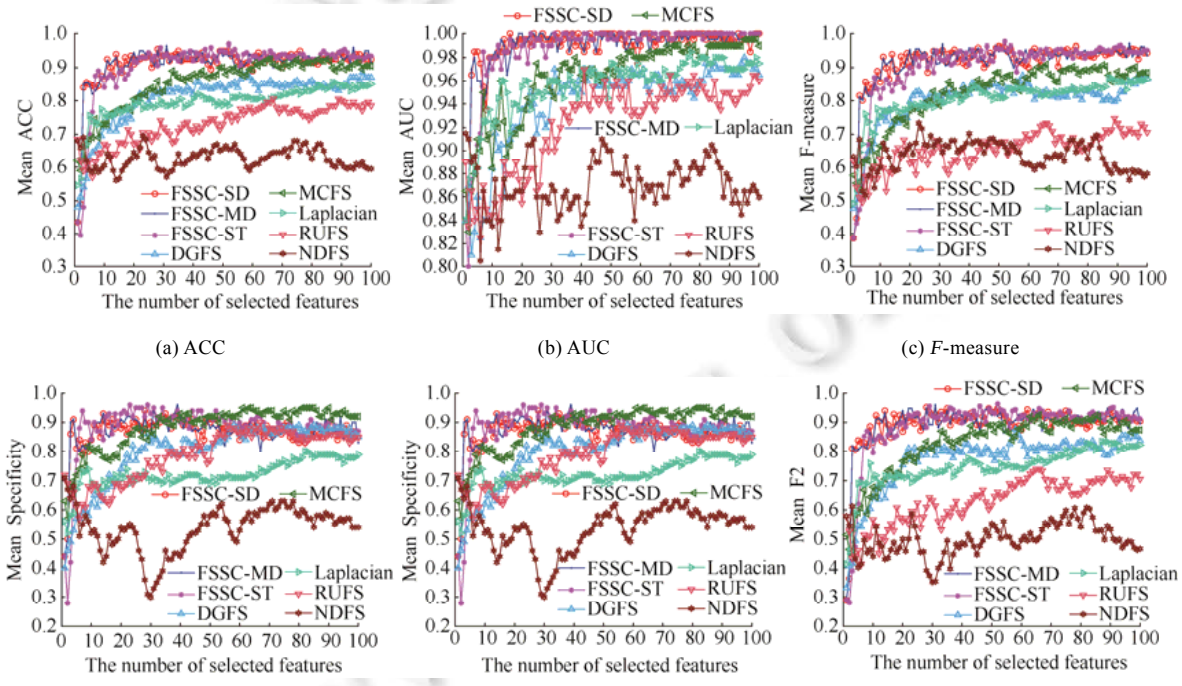


(a) ACC (b) AUC (c) *F*-measure

(d) Sensitivity (e) Specificity (f) *F*2

Fig.3 The average indexes of KNN classifier of each algorithm for different feature subsets on ALL1 dataset

图 3 各算法在 ALL1 数据集上对应不同特征子集的 KNN 分类器平均指标值

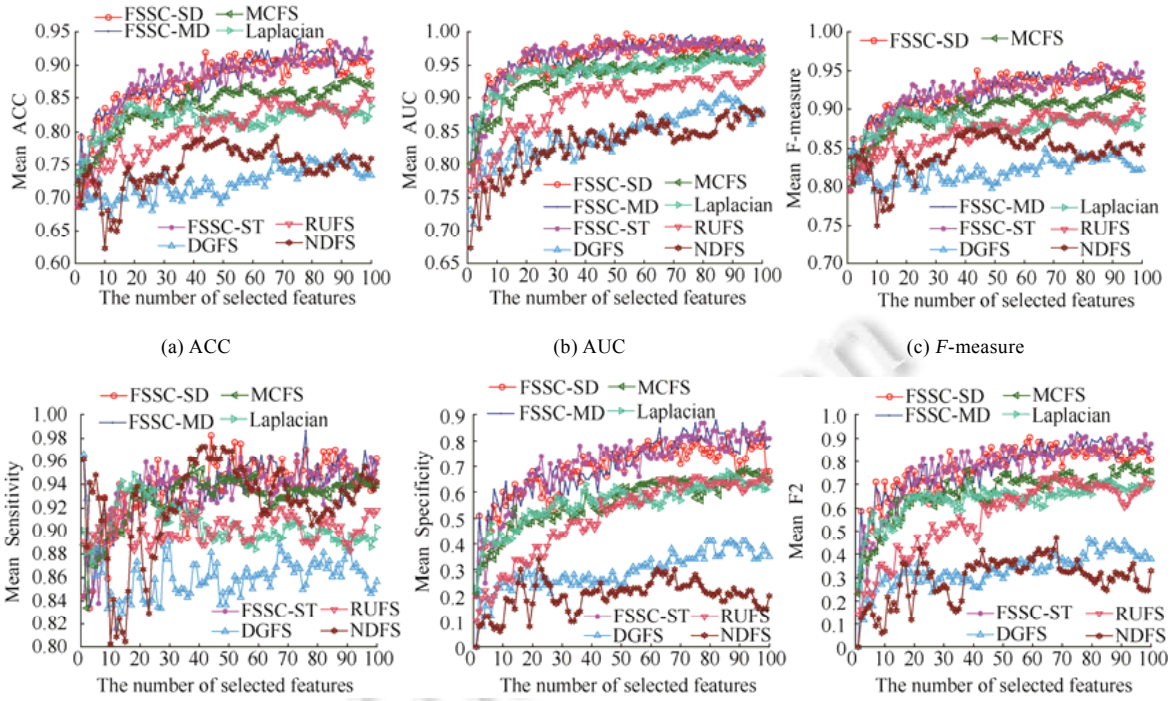


(a) ACC (b) AUC (c) *F*-measure

(d) Sensitivity (e) Specificity (f) *F*2

Fig.4 The average indexes of KNN classifier of each algorithm for different feature subsets on Carcinoma dataset

图 4 各算法在 Carcinoma 数据集上对应不同特征子集的 KNN 分类器平均指标值



(a) ACC (b) AUC (c) *F*-measure

(d) Sensitivity (e) Specificity (f) *F*2

Fig.5 The average indexes of KNN classifier of each algorithm for different feature subsets on DLBCL-Tumor dataset

图 5 各算法在 DLBCL-Tumor 数据集上对应不同特征子集的 KNN 分类器平均指标值

图 4 所示实验结果显示,本文提出的 FSSC-ST、FSSC-SD 和 FSSC-MD 算法选择的基因子集的分类性能优于 DGFS、Laplacian、RUFS 与 NDFS 算法选择的基因子集的分类性能.DGFS、MCFS、Laplacian 算法选择基因子集的分类性能居中,RUFS 和 NDFS 算法选择的基因子集的各项指标值最低.当选择的特征数较多时,MCFS 算法选择的基因子集的 KNN 分类器的 Sensitivity 和 Specifity 指标上略有超过 FSSC-ST、FSSC-SD 和 FSSC-MD 算法.因此,Carcinoma 数据集的实验结果揭示,本文所提算法 FSSC-SD、FSSC-MD 和 FSSC-ST 均能选择出区分能力好且包含特征数少的特征子集.

图 5 所示实验结果显示,本文提出的 FSSC-ST、FSSC-SD 和 FSSC-MD 算法选择的基因子集对应分类器的 ACC、AUC、*F*-measure、Specificity 和 *F*2 指标非常好,优于对比算法 DGFS、MCFS、Laplacian、RUFS 与 NDFS.在基因子集规模大于 50 时,本文提出的 FSSC-ST、FSSC-SD 和 FSSC-MD 算法选择的基因子集的 Sensitivity 指标优于其他对比算法.因此,本文提出的 FSSC-SD、FSSC-MD 和 FSSC-ST 算法均能选择到区分能力较好的特征子集.

综合图 2~图 5 的实验结果来看,本文提出的 FSSC-SD、FSSC-MD、FSSC-ST 算法均能选择出类别区分能力很好的特征子集,优于其他对比算法.

3.2.2 各算法最优值比较

为了验证本文算法的整体性能,比较各算法在 10 个癌症基因数据 5 次 10 折交叉验证选择的特征子集对应分类器的各指标平均结果的最优值,采用各算法在 10 个数据集实验结果的 win/draw/loss 来评价其性能.表 2、表 3 分别展示了各算法在 10 个数据集上所选特征子集对应 KNN、SVM 分类器的最优平均分类准确率 ACC、AUC、*F*-measure、Sensitivity、Specificity 和 *F*2 值的 win/draw/loss 比较.表 4 给出了各算法在表 1 的 10 个数据集选择的基因子集的 KNN 和 SVM 分类器各指标值的最优值平均的 win/draw/loss 结果比较.表中加粗和下

划线表示本文提出的 FSSC-SD、FSSC-MD 和 FSSC-ST 算法优于其他对比算法的结果.表中加粗加红和下划线表示本文提出的 FSSC-SD、FSSC-MD 和 FSSC-ST 谱特征选择算法之间比较,win 大于 loss 的结果.

Table 2 The highest average index comparison of KNN of selected gene subsets on 10 datasets (win/draw/loss)

表 2 各算法在 10 个数据集所选基因子集的 KNN 分类器的最高平均分类性能指标比较(win/draw/loss)

Indexes	Algorithms	FSSC-SD	FSSC-MD	FSSC-ST	DGFS	MCFS	Laplacian	RUFS	NDFS
ACC	FSSC-SD	–	**6/1/3**	**6/1/3**	**9/0/1**	**10/0/0**	**9/0/1**	**10/0/0**	**10/0/0**
	FSSC-MD	3/1/6	–	4/2/4	**9/0/1**	**10/0/0**	**9/0/1**	**10/0/0**	**10/0/0**
	FSSC-ST	3/1/6	4/2/4	–	**9/0/1**	**10/0/0**	**9/0/1**	**10/0/0**	**10/0/0**
AUC	FSSC-SD	–	3/2/5	**6/2/2**	**8/0/2**	**9/0/1**	**8/1/1**	**9/0/1**	**9/1/0**
	FSSC-MD	**5/2/3**	–	4/2/4	**9/0/1**	**9/0/1**	**9/1/0**	**9/0/1**	**9/1/0**
	FSSC-ST	2/2/6	4/2/4	–	**9/0/1**	**9/0/1**	**8/1/1**	**9/0/1**	**9/1/0**
F-measure	FSSC-SD	–	**7/1/2**	4/1/5	**9/0/1**	**10/0/0**	**9/0/1**	**10/0/0**	**9/0/1**
	FSSC-MD	2/1/7	–	4/1/5	**9/0/1**	**10/0/0**	**9/0/1**	**10/0/0**	**9/0/1**
	FSSC-ST	**5/1/4**	**5/1/4**	–	**9/0/1**	**10/0/0**	**9/0/1**	**10/0/0**	**10/0/0**
Sensitivity	FSSC-SD	–	**5/2/3**	**7/2/1**	**9/0/1**	**10/0/0**	**8/1/1**	**10/0/0**	**8/1/1**
	FSSC-MD	3/2/5	–	**5/2/3**	**8/0/2**	**10/0/0**	**8/1/1**	**10/0/0**	**8/1/1**
	FSSC-ST	1/2/7	3/2/5	–	**8/0/2**	**10/0/0**	**8/1/1**	**10/0/0**	**7/1/2**
Specificity	FSSC-SD	–	**7/1/2**	**5/1/4**	**10/0/0**	**9/0/1**	**9/0/1**	**9/0/1**	**10/0/0**
	FSSC-MD	2/1/7	–	3/2/5	**10/0/0**	**8/0/2**	**9/0/1**	**7/0/3**	**10/0/0**
	FSSC-ST	4/1/5	**5/2/3**	–	**10/0/0**	**10/0/0**	**9/0/1**	**10/0/0**	**10/0/0**
*F*2	FSSC-SD	–	4/1/5	4/1/5	**10/0/0**	**10/0/0**	**9/0/1**	**10/0/0**	**10/0/0**
	FSSC-MD	**5/1/4**	–	**5/1/4**	**10/0/0**	**10/0/0**	**9/0/1**	**10/0/0**	**10/0/0**
	FSSC-ST	**5/1/4**	4/1/5	–	**10/0/0**	**10/0/0**	**9/0/1**	**10/0/0**	**10/0/0**

表 2 所示实验结果揭示,本文提出的 FSSC-SD、FSSC-MD 和 FSSC-ST 谱特征选择算法选择的基因子集的 KNN 分类器的分类性能绝对地优于对比算法 DGFS、MCFS、Laplacian、RUFS 和 NDFS 选择的基因子集的分类能力.所提出的 FSSC-SD、FSSC-MD 和 FSSC-ST 算法相比,FSSC-SD 算法选择的基因子集的分类性能最优,在 ACC、AUC、Sensitivity 和 Specificity 这 4 个指标上优于 FSSC-ST 算法,在 ACC、*F*-measure、Sensitivity 和 Specificity 这 4 个指标上优于 FSSC-MD 算法.所提出的 FSSC-MD 与 FSSC-ST 算法选择的基因子集的分类能力相当.在 *F*2 指标上,提出的 3 种谱特征选择算法 FSSC-SD、FSSC-MD 和 FSSC-ST 选择的基因子集的分类能力相当,FSSC-MD 略优于 FSSC-SD 和 FSSC-ST 算法.

Table 3 The highest average index comparison of SVM of selected gene subsets on 10 datasets (win/draw/loss)

表 3 各算法在 10 个数据集所选基因子集的 SVM 分类器的最高平均分类性能指标比较(win/draw/loss)

Indexes	Algorithms	FSSC-SD	FSSC-MD	FSSC-ST	DGFS	MCFS	Laplacian	RUFS	NDFS
ACC	FSSC-SD	–	4/1/5	3/1/6	**10/0/0**	**9/1/0**	**10/0/0**	**10/0/0**	**10/0/0**
	FSSC-MD	**5/1/4**	–	4/2/4	**10/0/0**	**10/0/0**	**10/0/0**	**10/0/0**	**10/0/0**
	FSSC-ST	**6/1/3**	4/2/4	–	**10/0/0**	**10/0/0**	**10/0/0**	**10/0/0**	**10/0/0**
AUC	FSSC-SD	–	**6/2/2**	**6/2/2**	**9/0/1**	**8/1/1**	**9/1/0**	**9/0/1**	**9/1/0**
	FSSC-MD	2/2/6	–	**6/2/2**	**9/0/1**	**8/1/1**	**9/1/0**	**9/0/1**	**9/1/0**
	FSSC-ST	2/2/6	2/2/6	–	**9/0/1**	**8/1/1**	**8/1/1**	**9/0/1**	**8/1/1**
F-measure	FSSC-SD	–	4/1/5	**5/1/4**	**10/0/0**	**9/0/1**	**9/0/1**	**10/0/0**	**9/0/1**
	FSSC-MD	**5/1/4**	–	4/2/4	**9/1/0**	**9/0/1**	**8/1/1**	**9/1/0**	**8/1/1**
	FSSC-ST	4/1/5	4/2/4	–	**9/1/0**	**9/0/1**	**8/1/1**	**9/1/0**	**8/1/1**
Sensitivity	FSSC-SD	–	**3/7/0**	**3/5/2**	**6/4/0**	**5/5/0**	**5/4/1**	**6/4/0**	**5/4/1**
	FSSC-MD	0/7/3	–	**3/5/2**	**6/4/0**	**5/5/0**	**5/4/1**	**6/4/0**	**5/4/1**
	FSSC-ST	2/5/3	2/5/3	–	**6/4/0**	**5/5/0**	**5/4/1**	**6/4/0**	**5/4/1**
Specificity	FSSC-SD	–	2/4/4	2/4/4	**7/0/3**	**8/0/2**	**7/0/3**	**8/0/2**	**7/0/3**
	FSSC-MD	**4/4/2**	–	3/4/3	**7/0/3**	**8/0/2**	**7/0/3**	**8/0/2**	**7/0/3**
	FSSC-ST	**4/4/2**	3/4/3	–	**8/0/2**	**8/0/2**	**7/0/3**	**7/1/2**	**7/0/3**
*F*2	FSSC-SD	–	4/1/5	**5/1/4**	**10/0/0**	**10/0/0**	**9/0/1**	**10/0/0**	**9/0/1**
	FSSC-MD	**5/1/4**	–	**6/1/3**	**10/0/0**	**10/0/0**	**9/0/1**	**10/0/0**	**9/0/1**
	FSSC-ST	4/1/5	3/1/6	–	**10/0/0**	**10/0/0**	**9/0/1**	**10/0/0**	**9/0/1**

表 3 关于各算法选择的基因子集的 SVM 分类器的各项最优平均值比较显示:本文提出的 FSSC-ST、FSSC-MD 和 FSSC-ST 算法绝对地优于对比算法 DGFS、MCFS、Laplacian、RUFS 和 NDFS.所提出的 FSSC-SD、

FSSC-MD 和 FSSC-ST 这 3 种特征选择算法选择的基因子集对应 SVM 分类器的平均 AUC 最高值比较显示,FSSC-SD 选择的基因子集的性能最好,其次是 FSSC-MD 算法,FSSC-ST 位居第三.所提出的 3 种谱特征选择算法选择的基因子集对应 SVM 分类器的 ACC 比较显示,FSSC-ST 选择的基因子集的分类性能最好,然后依次是 FSSC-MD 和 FSSC-SD.*F*-measure 指标比较显示,FSSC-MD 算法的性能略优于 FSSC-SD 和 FSSC-ST 算法.Sensitivity 比较显示,提出的 3 种谱特征选择算法 FSSC_SD、FSSC-MD 与 FSSC-ST 选择的基因子集的分类性能基本相当,FSSC-SD 略优于 FSSC-MD 和 FSSC-ST,FSSC-MD 略优于 FSSC-ST.Specificity 最高均值比较显示,提出的 FSSC-ST 与 FSSC-MD 算法略优于提出的 FSSC-SD 算法,FSSC-ST 与 FSSC-MD 性能相当.各算法的基因子集对应 SVM 分类器的 *F*2 值显示,FSSC-MD 算法最优,然后依次是 FSSC-SD 算法和 FSSC-ST 算法.

由表 3 所示各算法选择的基因子集对应 SVM 分类器的实验结果分析得出:本文提出的 3 种无监督特征选择算法 FSSC-SD、FSSC-MD 和 FSSC-ST 绝对地优于对比算法 DGFS、MCFS、Laplacian、RUFS 和 NDFS.在所提出的 3 种算法之间,多数情况下的性能指标值比较基本持平,没有一种算法绝对地优于其他两种算法.

Table 4 The mean highest average index of KNN and SVM for the selected gene subsets on 10 datasets (win/draw/loss)

表 4 各算法在 10 个数据集所选基因子集的 SVM 与 KNN 分类器的最高平均分类指标的均值比较(win/draw/loss)

Index	Algorithms	FSSC-SD	FSSC-MD	FSSC-ST	DGFS	MCFS	Laplacian	RUFS	NDFS
ACC	FSSC-SD	–	7/1/2	4/1/5	10/0/0	10/0/0	10/0/0	10/0/0	10/0/0
	FSSC-MD	2/1/7	–	4/1/5	10/0/0	10/0/0	10/0/0	10/0/0	10/0/0
	FSSC-ST	5/1/4	5/1/4	–	10/0/0	10/0/0	10/0/0	10/0/0	10/0/0
AUC	FSSC-SD	–	6/2/2	6/2/2	9/0/1	9/0/1	9/1/0	9/0/1	9/1/0
	FSSC-MD	2/2/6	–	6/2/2	9/0/1	9/0/1	9/1/0	9/0/1	9/1/0
	FSSC-ST	2/2/6	2/2/6	–	9/0/1	9/0/1	9/1/0	9/0/1	9/1/0
F-measure	FSSC-SD	–	7/1/2	5/1/4	9/0/1	10/0/0	9/0/1	10/0/0	10/0/0
	FSSC-MD	2/1/7	–	4/1/5	9/0/1	10/0/0	9/0/1	10/0/0	9/0/1
	FSSC-ST	4/1/5	5/1/4	–	9/0/1	10/0/0	9/0/1	10/0/0	10/0/0
Sensitivity	FSSC-SD	–	5/2/3	7/2/1	9/0/1	10/0/0	8/1/1	10/0/0	8/1/1
	FSSC-MD	3/2/5	–	4/2/4	9/0/1	10/0/0	8/1/1	10/0/0	8/1/1
	FSSC-ST	1/2/7	4/2/4	–	9/0/1	10/0/0	8/1/1	10/0/0	7/1/2
Specifity	FSSC-SD	–	6/1/3	5/1/4	10/0/0	7/1/2	9/0/1	8/0/2	9/0/1
	FSSC-MD	3/1/6	–	2/4/4	10/0/0	8/0/2	10/0/0	8/0/2	10/0/0
	FSSC-ST	4/1/5	4/4/2	–	10/0/0	8/0/2	10/0/0	8/0/2	10/0/0
*F*2	FSSC-SD	–	5/1/4	4/1/5	10/0/0	10/0/0	10/0/0	10/0/0	10/0/0
	FSSC-MD	4/1/5	–	6/1/3	10/0/0	10/0/0	10/0/0	10/0/0	10/0/0
	FSSC-ST	5/1/4	3/1/6	–	10/0/0	10/0/0	10/0/0	10/0/0	10/0/0

表 4 给出关于提出的 FSSC-SD、FSSC-MD 和 FSSC-ST 谱特征选择算法在表 1 给出的 10 个基因数据集选择的基因子集对应 KNN 和 SVM 分类器相应最优指标的均值比较显示:所提出的 3 种谱特征选择算法所选基因子集的分类性能绝对地优于对比算法 DGFS、MCFS、Laplacian、RUFS 和 NDFS.所提出的算法 FSSC-SD、FSSC-MD 与 FSSC-ST 相比,FSSC-SD 算法选择的基因子集的分类能力在多数情况下优于所提出的 FSSC-MD 和 FSSC-ST 算法,FSSC-MD 在 AUC 和 *F*2 指标绝对地优于 FSSC-ST 算法,FSSC-ST 只在 ACC 和 *F*2 指标略优于 FSSC-SD 或 FSSC-MD.

因此,综合表 2～表 4 的实验结果可以得出:所提出的 FSSC-SD、FSSC-MD 和 FSSC-ST 谱特征选择算法绝对地优于对比算法 DGFS、MCFS、Laplacian、RUFS 和 NDFS.提出的 3 种谱特征选择算法彼此相比,FSSC-SD 选择的特征子集的分类能力最强,其次是 FSSC-MD 和 FSSC-ST 算法.

3.2.3 特征子集规模比较

前两小节展示了所提出的 3 种无监督特征选择算法选择的基因子集的分类性能比较,本小节将比较各算法选择的特征子集的规模,即选择的基因数.图 6 展示了各算法选择的基因子集对应 KNN 和 SVM 分类器的最优指标值对应基因子集包含的平均基因数.图中颜色越偏于 ColorBar 底部颜色,表示基因数越少(特征子集规模越小),反之,颜色越偏于 ColorBar 顶部的颜色,特征子集包含的基因数就越多.

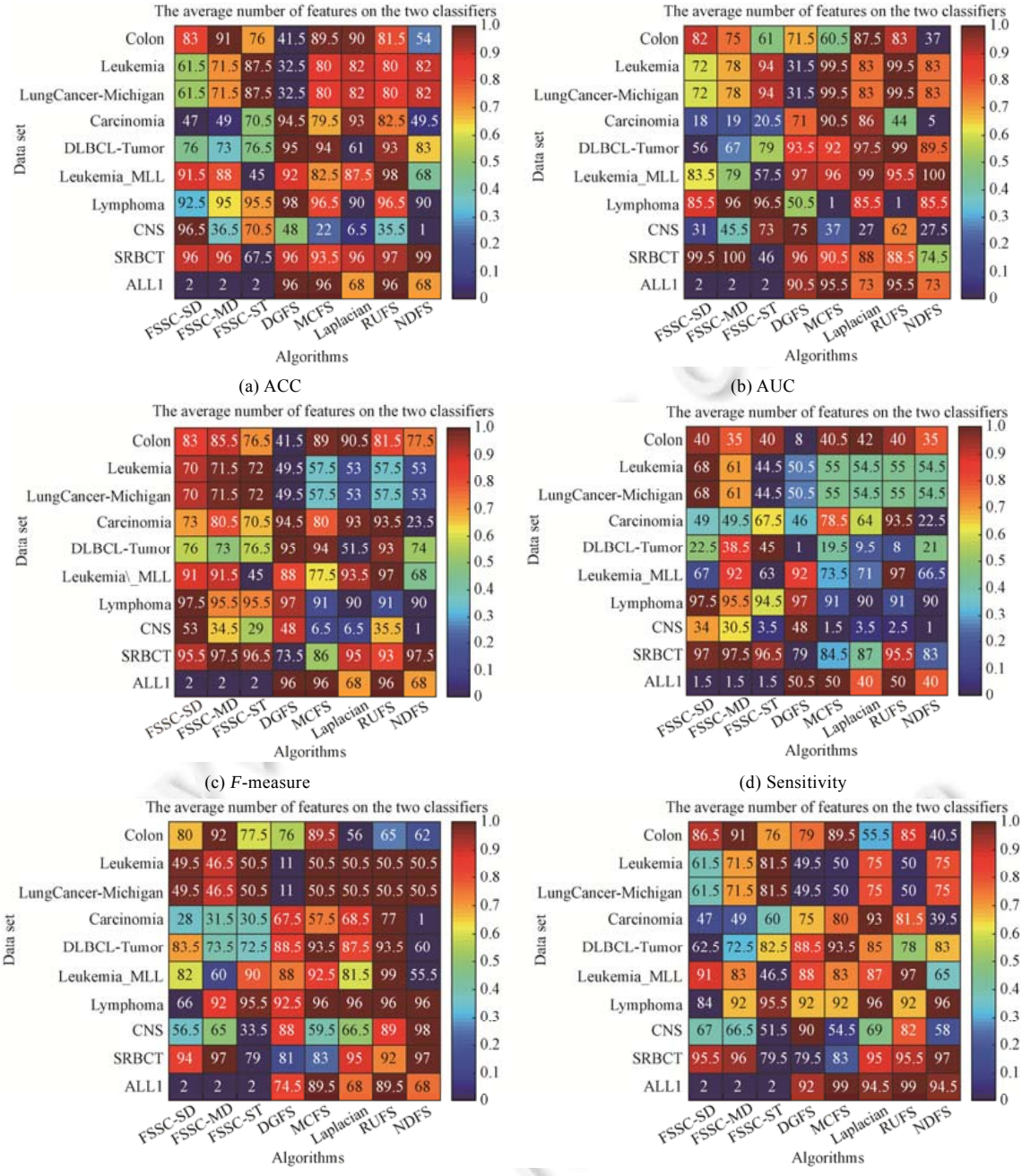


Fig.6 The mean feature number of selected gene subsets with the best average index of SVM and KNN by each algorithm for each dataset

图 6 各算法对各数据集选择的基因子集对应 SVM 和 KNN 分类器的最优指标值对应特征子集的平均特征数

图 6 所示特征子集规模比较显示,本文提出的 FSSC-SD、FSSC-MD 和 FSSC-ST 算法无论采用哪种评价指标,均能发现在 ALL1 数据集的最具分类能力且规模最小的基因子集.除了 *F*-measure 指标,在 ACC、AUC、*F*2、Sensitivity 和 Specificity 指标上,所提出的 FSSC-SD、FSSC-MD 和 FSSC-ST 算法在各数据集均能选择到规模

相对较小且具有很好分类能力的基因子集.分析其原因是,*F*-measure 指标过分强调了对于正类的识别能力,而忽略了选择的基因子集对于负类的识别能力.

综合对图 6 所示各算法选择的特征子集规模的分析可知,所提出的 3 种谱特征选择算法 FSSC-SD、FSSC-MD 和 FSSC-ST 能够选择到分类性能好且规模不大的特征子集.

4 各算法统计重要度分析

为了检验本文提出的 FSSC-SD、FSSC-MD 和 FSSC-ST 算法与对比算法 DGFS[23]、MCFS[19]、Laplacian[18]、RUFS[25]和 NDFS[26]是否具有统计显著性,本节采用 Friedman 检验方法来检验各特征选择算法间的差异[36,37],在用 Friedman 检测到算法间的显著性差异后,采用多重比较检验(multiple comparison test)作为事后检验,以发现各对特征选择算法之间的显著性差异.我们依据各算法在 10 个癌症基因数据 5 次 10 折交叉验证选择的特征子集对应 KNN 分类器的 ACC、AUC、*F*-measure、Sensitivity、Specificity 和 *F*2 各指标平均结果的最优值,在 α=0.05 时,进行 Friedman 检测.

6 种指标下的 Friedman 检测显示各特征选择算法之间存在显著差异.基于 KNN 分类器的预测准确率 ACC 的 χ^2=60.9034,*df*=7,*p*=9.9624e-11;AUC 的 χ^2=59.4230,*df*=7,*p*=1.9679e-10;*F*-measure 的 χ^2=58.7260,*df*=7,*p*=2.7102e-10;Sensitivity 的χ^2=44.7006,*df*=7,*p*=1.5633e-07;Specificity 的χ^2=23.7982,*df*=7,*p*=0.0012;*F*2 的χ^2=62.1587,*df*=7,*p*=5.5873e-11. *p*<0.05 对所有评价指标均成立,因此得出结论:各算法间存在显著差异.

表 5～表 10 展示了可信水平为 0.95 时,每一对特征选择算法进行多重比较检验的结果.各表上三角表示各对算法间的平均等级差,下三角表示各算法对之间的统计重要性,以*表示相应算法之间统计重要性显著.

表 5～表 10 所示的统计重要性检测结果显示:本文提出的 FSSC-ST 算法与所有对比特征选择算法 DGFS、MCFS、Laplacian、RUFS 和 NDFS 之间均存在显著性差异.对于所提出的 FSSC-SD 和 FSSC-MD 算法,当使用所选基因子集对应 KNN 分类器的 ACC、AUC、*F*-measure、Sensitivity 和 *F*2 指标时,与所有对比特征选择算法 DGFS、MCFS、Laplacian、RUFS 和 NDFS 之间均存在显著不同;当使用 Specificity 指标时,与 MCFS 算法没有显著区别,但与其他对比算法 DGFS、Laplacian、RUFS 和 NDFS 均存在显著性不同.算法 MCFS 与算法 Laplacian、RUFS 和 NDFS 在所有指标下均存在统计显著性不同.对比算法 NDFS 和 RUFS 在除了 Sensitivity 之外的其他指标上均存在显著性不同.

Table 5 Paired rank comparison of 8 feature selection algorithms in ACC of KNN predictive model

表 5 8 种特征选择算法依据特征子集对应 KNN 预测模型的 ACC 等级比较

Algorithm	FSSC-SD	FSSC-MD	FSSC-ST	DGFS	MCFS	Laplacian	RUFS	NDFS
FSSC-SD		–0.150 0	0.050 0	5.300 0	2.200 0	4.200 0	3.150 0	4.850 0
FSSC-MD			0.200 0	5.450 0	2.350 0	4.350 0	3.300 0	5.000 0
FSSC-ST				5.250 0	2.150 0	4.150 0	3.100 0	4.800 0
DGFS	*	*	*		–3.100 0	–1.100 0	–2.150 0	–0.450 0
MCFS	*	*	*			2.000 0	0.950 0	2.650 0
Laplacian	*	*	*		*		–1.050 0	0.650 0
RUFS	*	*	*		*			1.700 0
NDFS	*	*	*		*		*	

Table 6 Paired rank comparison of 8 feature selection algorithms in AUC of KNN predictive model

表 6 8 种特征选择算法依据对应特征子集的 KNN 预测模型的 AUC 等级比较

Algorithm	FSSC-SD	FSSC-MD	FSSC-ST	DGFS	MCFS	Laplacian	RUFS	NDFS
FSSC-SD		0	0.600 0	5.350 0	2.300 0	4.250 0	3.750 0	4.950 0
FSSC-MD			0.600 0	5.350 0	2.300 0	4.250 0	3.750 0	4.950 0
FSSC-ST				4.750 0	1.700 0	3.650 0	3.150 0	4.350 0
DGFS	*	*	*		–3.050 0	–1.100 0	–1.600 0	–0.400 0
MCFS	*	*	*			1.950 0	1.450 0	2.650 0
Laplacian	*	*	*		*		–0.500 0	0.700 0
RUFS	*	*	*		*			1.200 0
NDFS	*	*	*		*		*	

Table 7 Paired rank comparison of 8 feature selection algorithms in *F*-measure of KNN predictive model

表 7 8 种特征选择算法依据对应特征子集的 KNN 预测模型的 *F*-measure 等级比较

Algorithm	FSSC-SD	FSSC-MD	FSSC-ST	DGFS	MCFS	Laplacian	RUFS	NDFS
FSSC-SD		–0.300 0	–0.200 0	5.100 0	2.100 0	4.100 0	3.100 0	4.500 0
FSSC-MD			0.100 0	5.400 0	2.400 0	4.400 0	3.400 0	4.800 0
FSSC-ST				5.300 0	2.300 0	4.300 0	3.300 0	4.700 0
DGFS	*	*	*		–3.000 0	–1.000 0	–2.000 0	–0.600 0
MCFS	*	*	*			2.000 0	1.000 0	2.400 0
Laplacian	*	*	*		*		–1.000 0	0.400 0
RUFS	*	*	*		*			1.400 0
NDFS	*	*	*		*		*	

Table 8 Paired rank comparison of 8 feature selection algorithms in Sensitivity of KNN predictive model

表 8 8 种特征选择算法依据对应特征子集的 KNN 预测模型的 Sensitivity 等级比较

Algorithm	FSSC-SD	FSSC-MD	FSSC-ST	DGFS	MCFS	Laplacian	RUFS	NDFS
FSSC-SD		–0.200 0	0.900 0	4.250 0	2.100 0	3.450 0	3.350 0	3.350 0
FSSC-MD			1.100 0	4.450 0	2.300 0	3.650 0	3.550 0	3.550 0
FSSC-ST				3.350 0	1.200 0	2.550 0	2.450 0	2.450 0
DGFS	*	*	*		2.150 0	–0.800 0	–0.900 0	–0.900 0
MCFS	*	*	*			1.350 0	1.250 0	1.250 0
Laplacian	*	*	*		*		–0.100 0	–0.100 0
RUFS	*	*	*		*			0
NDFS	*	*	*		*			

Table 9 Paired rank comparison of 8 feature selection algorithms in Specificity of KNN predictive model

表 9 8 种特征选择算法依据对应特征子集的 KNN 预测模型的 Specificity 等级比较

Algorithm	FSSC-SD	FSSC-MD	FSSC-ST	DGFS	MCFS	Laplacian	RUFS	NDFS
FSSC-SD		0.200 0	0.100 0	3.150 0	0.550 0	2.300 0	1.550 0	3.150 0
FSSC-MD			–0.300 0	2.950 0	0.350 0	2.100 0	1.350 0	2.950 0
FSSC-ST				3.250 0	0.650 0	2.400 0	1.650 0	3.250 0
DGFS	*	*	*		–2.600 0	–0.850 0	–1.600 0	0
MCFS			*			1.750 0	1.000 0	2.600 0
Laplacian	*	*	*		*		–0.750 0	0.850 0
RUFS	*	*	*		*			1.600 0
NDFS	*	*	*		*		*	

Table 10 Paired rank comparison of 8 feature selection algorithms in *F*2 of KNN predictive model

表 10 8 种特征选择算法依据对应特征子集的 KNN 预测模型的 *F*2 等级比较

Algorithm	FSSC-SD	FSSC-MD	FSSC-ST	DGFS	MCFS	Laplacian	RUFS	NDFS
FSSC-SD		0.900 0	0.300 0	5.200 0	2.100 0	3.900 0	3.100 0	4.700 0
FSSC-MD	*		1.200 0	6.100 0	3.000 0	4.800 0	4.000 0	5.600 0
FSSC-ST				4.900 0	1.800 0	3.600 0	2.800 0	4.400 0
DGFS	*	*	*		–3.100 0	–1.300 0	–2.100 0	–0.500 0
MCFS	*	*	*			1.800 0	1.000 0	2.600 0
Laplacian	*	*	*		*		0.800 0	0.800 0
RUFS	*	*	*		*	*		1.600 0
NDFS	*	*	*		*	*	*	

表 10 所示统计分析结果显示:当使用 *F*2 指标时,本文提出的 3 种谱特征选择算法 FSSC-SD、FSSC-MD 和 FSSC-ST 不仅与所有对比算法 DGFS、MCFS、Laplacian、RUFS 和 NDFS 存在显著性不同,所提出的 FSSC-SD 和 FSSC-MD 算法之间也存在统计显著性差异.另外,Laplacian 和 RUFS、NDFS 算法之间也存在显著性不同.

由以上统计重要性分析可见,所提出的谱特征选择算法 FSSC-SD、FSSC-MD 和 FSSC-ST 是非常有效的基因选择算法,能够选择到分类性能显著不同于对比算法的基因子集.

5 结 论

针对癌症基因数据的特征选择问题,提出基于谱聚类的无监督特征选择算法 FSSC-SD、FSSC-MD 和 FSSC-ST 对所有特征进行谱聚类,将相似性较高(强冗余性)的特征聚成一类,从各类簇选择代表特征构成特征子集.提出特征区分度、特征独立性、特征重要度概念,定义特征区分度为其标准差,定义特征独立性为其与簇内其他

区分度更高特征的Pearson相关系数和的倒数,对区分度最大特征,定义其独立性为其与所在特征簇最不相关特征的 Pearson 相关系数绝对值的倒数,定义特征重要性为其区分度与独立性之积.10 个基因数据集的实验测试结果及各算法的统计显著性检测结果表明,所提出的无监督谱特征选择算法 FSSC-SD、FSSC-MD 和 FSSC-ST 均能选择到不仅具有强分类能力,且包含基因数较少的特征子集,其中,FSSC-SD 算法选择的特征子集的分类能力最优.所提出的谱特征选择算法 FSSC-SD、FSSC-MD 和 FSSC-ST 与对比算法 DGFS、MCFS、Laplacian、RUFS、NDFS 之间存在显著性差异.

References:

[1] Derisi JL, Iyer VR, Brown PO. Exploring the metabolic and genetic control of gene expression on a genomic scale. Science, 1997,278(5338):680–686.

[2] Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science, 1999,286(5439):531–537.

[3] Khan J, Wei JS, Ringner M, Saal LH, Ladanyi M, Westermann F, Berthold F, Schwab M, Antonescu CR, Peterson C. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nature Medicine, 2001,7(6): 673.

[4] Li YX, Li JG, Ruan XG. Study of informative gene selection for tissue classification based on tumor gene expression profiles. Chinese Journal of Computers, 2006,29(2):324−330 (in Chinese with English abstract).

[5] Guyon I, Elisseeff A. An introduction to variable and feature selection. Journal of Machine Learning Research, 2003,3:1157–1182.

[6] Ding C, Peng HC. Minimum redundancy feature selection from microarray gene expression data. Journal of Bioinformatics and Computational Biology, 2005,3(2):185–205.

[7] Xie JY, Wang MZ, Hu QF. The differentially expressed gene selection algorithms for unbalanced gene datasets by maximize the area under ROC. Journal of Shaanxi Normal University (Natural Science Edition), 2017,45(1):13–22 (in Chinese with English abstract).

[8] Xie JY, Wang MZ, Zhou Y, Li, JY. Coordinating discernibility and independence scores of variables in a 2D space for efficient and accurate feature selection. In: Huang DS, Han K, Hussain A, eds. Proc. of the Int'l Conf. on Intelligent Computing 2016, Part III, Intelligent Computing Methodologies. LNAI 9773, Springer Int'l Publishing Switzerland, 2016. 116–127.

[9] Blum AL, Langley P. Selection of relevant features and examples in machine learning. Artificial Intelligence, 1997,97(1-2): 245–271.

[10] Kohavi R, John GH. Wrappers for feature subset selection. Artificial Intelligence, 1997,97(1-2):273–324.

[11] Lal TN, Chapelle O, Weston J, Elisseeff A. Embedded methods. In: Guyon I, Nikravesh M, Gunn S, Zadeh LA, eds. Proc. of the Feature Extraction, Foundation and Applications, Studies in Fuzziness and Soft Computing, 207. Berlin, Heidelberg: Springer-Verlag, 2006. 137–165.

[12] Kira K, Rendell LA. The feature selection problem: Traditional methods and a new algorithm. In: Kerber R, ed. Proc. of the 10th National Conf. on Artificial Intelligence. AAAI Press, 1992. 129–134.

[13] Peng HC, Long FH, Ding C. Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans. on Pattern Analysis & Machine Intelligence, 2005,(8):1226–1238.

[14] Hall MA. Correlation-based feature selection for machine learning [Ph.D. Thesis]. Hamilton: University of Waikato, 1999.

[15] Dash M, Liu H, Yao J. Dimensionality reduction of unsupervised data. In: Storms P, ed. Proc. of the 9th IEEE Int'l Conf. on Tools with Artificial Intelligence. IEEE, 1997. 532–539.

[16] Xu JL, Zhou YM, Chen L, Xu BW. An unsupervised feature selection approach based on mutual information. Journal of Computer Research and Development, 2012,49(2):372–382 (in Chinese with English abstract).

[17] Zhang L, Sun G, Guo J. Unsupervised feature selection method based on *K*-means clustering. Application Research of Computers, 2005,22(3):22–24 (in Chinese with English abstract).

[18] He XF, Cai DY, Niyogi P. Laplacian score for feature selection. In: Weiss Y, Schölkopf B, Platt J, eds. Advances in Neural Information Processing Systems (NIPS18). Cambridge: MIT Press, 2006. 507–514.

[19] Cai D, Zhang CY, He XF. Unsupervised feature selection for multi-cluster data. In: Wallace BC, Small K, Brodley CE, Trikalinos TA, eds. Proc. of the 16th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. Washington: ACM, 2010. 333–342.

[20] Wang LX, Jiang SY. Novel feature selection method based on feature clustering. Application Research of Computers, 2015, 32(5):1305–1308 (in Chinese with English abstract).

[21] Zheng Z, Liu H. Spectral feature selection for supervised and unsupervised learning. In: Ghahramani Z, ed. Proc. of the 24th Int'l Conf. on Machine Learning (ICML 2007). New York: ACM, 2007. 1151–1157.

[22] Xie JY, Qu YN, Wang MZ. Unsupervised feature selection algorithms based on density peaks. Journal of Nanjing University (Natural Sciences). 2016,52(4):735–745.

[23] He JR, Bi YZ, Ding LX, Li ZK, Wang SW. Unsupervised feature selection based on decision graph. Neural Computing and Applications, 2017,28(10):3047–3059.

[24] Xie JY, Fan W. Gene markers identification algorithm for detecting colon cancer patients. Pattern Recognition and Artificial Intelligence, 2017,52(4):1019–1029 (in Chinese with English abstract).

[25] Qian MJ, Zhai CX. Robust unsupervised feature selection. In: Rossi F, ed. Proc. of the 23rd Int'l Joint Conf. on Artificial Intelligence (IJCAI 2013). AAAI Press, 2013. 1621–1627.

[26] Li ZC, Yang Y, Liu J, Zhou XF, Lu HQ. Unsupervised feature selection using nonnegative spectral analysis. In: Hoffmann J, Selman B, eds. Proc. of the 26th AAAI Conf. on Artificial Intelligence (AAAI-12). Toronto: AAAI Press, 2012. 1026–1032.

[27] Hu MJ, Zheng LP, Tang L, Yang H, Fu W. Feature selection algorithm based on joint spectral clustering and neighborhood mutual information. Pattern Recognition and Artificial Intelligence, 2017,30(12):1121–1129 (in Chinese with English abstract).

[28] Jiang SY, Zheng Q, Zhang QS. Clustering-based feature selection. Acta Electronica Sinica, 2008,36(s1):157–160. (in Chinese with English abstract).

[29] Xie JY, Ding LJ. The true self-adaptive spectral clustering algorithms. Acta Electronica Sinica, 2019,47(5):1000–1008 (in Chinese with English abstract).

[30] Zelnik-Manor L, Perona P. Self-tuning spectral clustering. In: Saul LK, Weiss Y, Bottou L, eds. Advances in Neural Information Processing Systems (NIPS17). Cambridge MA: MIT Press, 2005. 1601–1608.

[31] Wang L, Bo LF, Jiao LC. Density-sensitive semi-supervised spectral clustering. Ruan Jian Xue Bao/Journal of Software, 2007, 18(10):2412–2422 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/18/2412.htm [doi: 10.1360/jos182412]

[32] Xie JY, Zhou Y. A new criterion for clustering algorithm. Journal of Shaanxi Normal University (Natural Science Edition), 2015,43(6):1–8 (in Chinese with English abstract).

[33] Luxburg UV. A tutorial on spectral clustering. Statistics and Computing, 2007,17(4):395–416.

[34] Chang CC, Lin CJ. LIBSVM: A library for support vector machines. ACM Trans. on Intelligent Systems and Technology (TIST), 2011,2(3):27.

[35] Xie JY, Wang MZ, Zhou Y, Gao HC, Xu SQ. Differentially expressed gene selection algorithms for unbalanced gene datasets. Chinese Journal of Computers, 2019,42(6):1232–1251 (in Chinese with English abstract).

[36] Borg A, Lavesson N, Boeva V. Comparison of clustering approaches for gene expression data. In: Jaeger M, Nielsen TD, Viappiani P, eds. Proc. of the 12th Scandinavian Conf. on Artificial Intelligence. Aalborg: IOS Press, 2013. 55–64.

[37] Xie JY, Gao HC, Xie WX, Liu XH, Grant PW. Robust clustering by detecting density peaks and assigning points based on fuzzy weighted *K*-nearest neighbors. Information Sciences, 2016,354:19–40.

附中文参考文献:

[4] 李颖新,李建更,阮晓钢.肿瘤基因表达谱分类特征基因选取问题及分析方法研究.计算机学报,2006,29(2):324−330.

[7] 谢娟英,王明钊,胡秋锋.最大化 ROC 曲线下面积的不平衡基因数据集差异表达基因选择算法.陕西师范大学学报(自然科学版),2017,45(1):13–22.

[16] 徐峻岭,周毓明,陈林,徐宝文.基于互信息的无监督特征选择.计算机研究与发展,2012,49(2):372–382.

[17] 张莉,孙钢,郭军.基于 *K*-均值聚类的无监督的特征选择方法.计算机应用研究,2005,22(3):22–24.

[20] 王连喜,蒋盛益.一种基于特征聚类的特征选择方法.计算机应用研究,2015,(5):1305–1308.

[22] 谢娟英,屈亚楠,王明钊.基于密度峰值的无监督特征选择算法.南京大学学报(自然科学),2016,52(4):735–745.

[24] 谢娟英,樊雯.结肠癌患者诊断的基因标志物识别算法.模式识别与人工智能,2017,52(4):1019–1029.

[27] 胡敏杰,郑荔平,唐莉,杨红,郑荔平,傅为.联合谱聚类与邻域互信息的特征选择算法.模式识别与人工智能,2017,30(12):1121–1129.

[28] 蒋盛益,郑琪,张倩生.基于聚类的特征选择方法.电子学报,2008,36(s1):157–160.

[29] 谢娟英,丁丽娟.完全自适应的谱聚类算法.电子学报,2019,47(5):1000–1008.

[31] 王玲,薄列峰,焦李成.密度敏感的半监督谱聚类.软件学报,2007,18(10):2412−2422. http://www.jos.org.cn/1000-9825/18/2412.htm [doi: 10.1360/jos182412]

[32] 谢娟英,周颖.一种新聚类评价指标.陕西师范大学学报(自然科学版),2015,43(6):1–8.

[35] 谢娟英,王明钊,周颖,高红超,许升全.非平衡基因数据的差异表达基因选择算法研究.计算机学报,2019,42(6):1232–1251.

谢娟英(1971－),女,陕西西安人,博士,教授,博士生导师,CCF 高级会员,主要研究领域为机器学习,数据挖掘,生物医学数据分析.

丁丽娟(1994－),女,硕士生,主要研究领域为机器学习,数据挖掘.

王明钊(1990－),男,博士生,主要研究领域为数据挖掘,生物信息学.