计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2018,27(3):149−155 [doi: 10.15888/j.cnki.csa.006243]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于改进 ReliefF 的无监督特征选择方法①

丁雪梅[1,2], 王汉军[1], 王炤光[3], 周心圆[4]

[1](中国科学院 沈阳计算技术研究所, 沈阳 110168)
[2](中国科学院大学, 北京 100049)
[3](国家电网公司东北分部, 沈阳 110180)
[4](吉林大学 计算机科学与技术学院, 长春 130000)
通讯作者: 丁雪梅, E-mail: dingxuemei15@mails.ucas.ac.cn

摘　要: 针对特征选择中存在数据缺乏类别信息的问题, 提出一种新型的基于改进 ReliefF 的无监督特征选择方法 UFS-IR. 由于 ReliefF 类算法存在小类样本抽样概率低、无法删除冗余特征的缺陷, 该方法以 DBSCAN 聚类算法指导分类, 通过改进抽样策略, 使用调整的余弦相似度度量特征间的相关性作为去冗余的凭据. 实验表明 UFS-IR 可以有效缩减数据维度的同时保证特征子集的最大相关最小冗余性, 具有很好的性能.

关键词: DBSCAN; ReliefF; 调整的余弦相似度; 无监督特征选择

引用格式: 丁雪梅,王汉军,王炤光,周心圆.基于改进 ReliefF 的无监督特征选择方法.计算机系统应用,2018,27(3):149–155. http://www.c-s-a.org.cn/1003-3254/6243.html

Unsupervised Feature Selection Method Based on Improved ReliefF

DING Xue-Mei[1,2], WANG Han-Jun[1], WANG Zhao-Guang[3], ZHOU Xin-Yuan[4]

[1](Shenyang Institute of Computing Technology, Chinese Academy of Science, Shenyang 110168, China)
[2](University of Chinese Academy of Sciences, Beijing 100049, China)
[3](State Grid Corporation Northeast Branch Corporation, Shenyang 110180, China)
[4](College of Computer Science and Technology, Jilin University, Changchun 130000, China)

Abstract: A novel method of unsupervised feature selection UFS-IR based on improved ReliefF is proposed to solve the problem of lack of category information in feature selection. As the ReliefF algorithm has a small sampling probability of small class samples, it cannot delete the defects of redundant features. This method uses DBSCAN clustering algorithm to guide the classification. By improving the sampling strategy, it uses the adjusted cosine similarity to measure the correlation between features as a de-redundancy credential. Experiments show that UFS-IR can effectively reduce the data dimension while ensuring the maximum correlation redundancy of the feature subset, and with good performance.

Key words: DBSCAN; ReliefF; adjusted cosine similarity; unsupervised feature selection

特征选择是指从原始特征集中选出一个使得评估标准达到最优的特征子集的过程[1]. 对于一个概念学习问题, 优秀的学习样本是训练分类器的关键, 样本中的不相关或冗余特征会使学习算法陷入混乱导致分类器过拟合[2], 影响训练性能. 因此有效的特征选择对于加速学习速度和提高概念质量具有重要作用.

特征选择根据是否包含类别信息分为有监督的特征选择和无监督的特征选择. 无监督特征选择的目的是根据一定的评判标准, 选择出一个足够简练又能够充分描绘原始特征集的重要特性同时保障数据集原生分类性的特征子集. 针对无监督特征选择, 已经提出了一些方法, 根据是否于后续的学习方法相关, 可分为过

① 基金项目: 国科控股企业技术创新引导基金 (2015XS0356)
收稿时间: 2017-06-14; 修改时间: 2017-06-30; 采用时间: 2017-07-08; csa 在线出版时间: 2018-02-09

滤式 (Filter) 和封装式 (Wrapper) 两种[3]. Filter 模型独立于学习算法, 利用所有训练数据的统计性能, 选取相关性评价准则进行特征评估. 文献[4]引入核空间距离测度, 通过计算两类样本点在核空间的距离度量相关性, 有效提高了线性不可分数据的特征选择能力. 文献[5]采用互信息来度量相关性, 以此达到无监督最小冗余最大相关 (UmRMR) 的特征选择的标准. 文献[6]使用拉普拉斯分值评价特征的重要程度, 以极大程度保留原始特征集和整体几何结构信息为原则选择分值较小的部分特征形成特征子集. Tabakhi S[7]以特征作为结点, 特征间余弦相似度作为边的权值建立图模型; 使用蚁群算法, 引入信息素的概念启发式搜索相似度最小的路径, 所历结点形成最终特征子集. 然而以上过滤式方法单纯以数据本身的统计信息或关系作为依据进行特征选择, 没有考虑特征对于数据原生性分类的影响. 特征选择知识理论中分类性能也可作为所选特征子集的度量标准. 在 Wrapper 模型中, 特征选择算法与学习算法耦合在一起, 利用学习算法的分类准确率评估特征子集. 针对无监督问题则是以特定聚类算法的聚类结果的质量来估量特征子集. 例如 Yuan 等人[8]对数据域进行特征聚类, 然后通过构建熵测度来揭示不同特征子集的最优值以此评估该特征子集的重要性. Zhu 等人[9]提出了一种子空间聚类指导特征选择的方法, 该方法维持并迭代更新一个特征选择矩阵, 矩阵的列向量反馈每个子空间聚类的代表性特征. 但是该类方法计算量大, 时间复杂度较高, 无法适用于大规模数据的特征选择.

ReliefF 是公认的效果较好的过滤式 (Filter) 特征评估方法[10], 不同于上述过滤式方法, ReliefF 利用特征对于分类的影响来评估特征权重, 同时对数据类型没有限制, 运行效率高, 能够应对大规模的数据. 针对其易忽略小类样本、不能削减冗余特征等不足, 本文提出了一种基于改进 ReliefF 的无监督特征选择算法 (UFS-IR), 继承 ReliefF 优点的同时特征选择结果更可靠、更准确, 获得符合 UmRMR 标准的特征子集.

1 无监督特征选择

无监督特征选择是指按照某一评价准则从原始高维特征集中筛选出部分特征形成最优特征子集, 使它对原始数据的自然分类效果的影响足够小或者没有.

定义已知包含 n 个特征 $X_1, X_2, \cdots, X_n$ 的数据集 D, 其中 $X_i=(S_1,S_2,S_3,\cdots,S_k)$, $k\in N^*$(非零自然数); 给定方法 *Method*, 按照某一准则 J 评价各个特征 X_i, 筛选出其中部分特征 $X_i, X_j, \cdots, X_m(1\le i, j, m\le n$ 且 $i, j, m\in N^*)$ 获取最终的特征子集$F\subsetneq D$.

2 改进 ReliefF 算法

2.1 ReliefF 算法

Relief 系列算法一种高效的过滤式特征选择方法, 最早于 1992 年由 Kira 和 Rendell 提出用于二分类问题的特征选择算法[11]. 1994 年 Kononenko 对 Relief 进行分析和扩展, 使得 ReliefF 能够用来处理噪声、不完整和多类数据集[12].

ReliefF 基于特征对各个类的近距离样本的区分能力, 赋予特征不同的权重来评估特征, 特征权值越大意味着它更有助于区分类别. 当特征与分类相关性极低时, 特征的权值将足够小接近 0; 特征权值计算结果可能出现负值, 表示同类近邻样本的距离比不同类近邻样本的距离大, 这也意味着该特征对于分类的影响是负面的.

对于样本集 Q, 每次从中随机选择一个样本 S, 然后 S 的同类样本集中寻找 k 个 S 的近邻样本 *NH*, 同时每个与 S 不同类别的样本集中各寻找 k 个近邻样本 *NM*. 迭代更新每个特征的权重 $\omega(x)$, 更新公式为:

$$\omega(x)=\omega(x)-\sum_{j=1}^{k} diff(X,S,NH_j)/(m*k)+\sum_{C\neq Class(S)}\left[\frac{P(C)}{1-P(Class(S))}*\sum_{j=1}^{k} diff(X,S,NM(C)_j)\right]/(m*k) \tag{1}$$

$$diff(X,S,S')=\begin{cases}(|S[X]-S'[X]|)/(\max(X)-\min(X)), & X\text{连续}\\ 0, & X\text{离散,且}S[X]=S'[X]\\ 1, & X\text{离散,且}S[X]\neq S'[X]\end{cases} \tag{2}$$

式中, m 表示迭代次数, NH_j 表示同类的第 j 个近邻样本, $NM(C)_j$ 表示不同类的 C 类样本的第 j 个近邻样本, $P(C)$ 表示第 C 类目标的概率, $Class(S)$ 表示样本 S 所属的类别, $diff(X, S, S')$ 表示样本 S 和 S'关于特征 X 的距离.

2.2 改进 ReliefF 算法

虽然 ReliefF 算法不限制数据类型为离散或是连续型, 能应对多类别的数据, 拥有较高的评估效率和有效性, 但它仍存在以下不足.

(1) 算法随机采样的次数 m 将影响最终得出的各个特征的权值.

(2) 随机采样的策略使得小类样本被选中的几率很低甚至不被选中, 如果忽略小类别样本对更新特征权重的影响, 那么最终的结果准确性和合理性是不可靠的.

(3) 随机采样可能存在重复抽样的情况, 重复抽样的样本将对特征权值的更新产生无意义的重复影响, 使得有限抽样次数内的有效抽样率降低, 从而降低了特征权值结果的有效性和可靠性.

(4) 算法得出的特征权值, 只能评估特征对分类的贡献值并不能帮助删除冗余特征.

本文针对以上不足的 (2)(3)(4) 对 ReliefF 算法加以改善. 在抽样的随机性不变的前提下, 控制每一类样本被抽样到的次数, 以各类样本占所有有效样本的比例 ξ 设置每一类样本被抽样的次数 ξm, 弥补 ReliefF 算法随机选样时小类别样本被选中概率低的不足. 同时控制每一次抽样的样本不是重复抽样, 总是对特征权值更新赋予新的影响, 充分发挥数据集在有限次数内对特征评估的效用. 针对 ReliefF 算法无法删除冗余特征, 引进调整的余弦相似度来衡量特征间的相关性, 结合特征评估结果去除相关特征对中的一个元素.

2.3 调整的余弦相似度

余弦相似度通过计算两个向量的夹角余弦值来评估他们的相似度. 余弦值的范围在[–1, 1]之间, 值越趋近于 1, 代表两个向量的方向越接近越相似; 越趋近于 –1, 他们的方向越相反; 接近于 0, 表示两个向量近乎于正交.

向量 $X=(X_1,X_2,\cdots,X_n)$ 和 $Y=(Y_1,Y_2,\cdots,Y_n)$ 的余弦相似度为:

$$\cos(X,Y)=\frac{\sum_{i=1}^{n}X_i*Y_i}{\sqrt{\sum_{i=1}^{n}X_i^2}*\sqrt{\sum_{i=1}^{n}Y_i^2}} \tag{3}$$

余弦相似度衡量的是空间向量的夹角, 仅仅从方向上区分向量的差异, 而对绝对数值不敏感, 向量数值的成比例缩放不改变余弦相似度计算结果. 余弦相似度对数值的不敏感导致了结果的误差, 调整的余弦相似度 ACS 修正了这种不合理性, 在向量所有维度上的数值 X_i 都减去一个均值$\bar{X}$[13], 也使得衡量向量的相似性转变为衡量相关性.

向量 $X=(X_1,X_2,\cdots,X_n)$ 和向量 $Y=(Y_1,Y_2,\cdots,Y_n)$ 的调整余弦相似度为:

$$ACS(X,Y)=\cos(X-\bar{X},Y-\bar{Y})=\frac{\sum_{i=1}^{n}(X_i-\bar{X})*(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}*\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}} \tag{4}$$

3 算法设计

3.1 基础算法

传统 ReliefF 算法的设计是针对于有监督的特征选择, 数据的类别信息是算法的主要支撑. 引入 ReliefF 算法进行无监督特征选择的首要任务就是获知数据类别. 聚类是将样本观测值, 数据项或特征向量划分成簇的无监督分类模式[14], 能够完成对无监督样本的分类. DBSCAN 是一种基于密度聚类算法的典型代表, 它能够把具有足够密度的区域划分为簇, 聚类速度快, 对聚类簇的形状毫无偏倚, 能够在具有噪声的空间数据库中发现任意形状的簇, 同时有效处理噪声数据防止对有效样本形成的评估结果产生干扰. 因此本文以 DBSCAN 聚类结果形成的簇定义数据的类别, 以此支撑 UFS-IR 算法.

DBSCAN 定义参数 *eps* 来表示每个对象的邻域半径, 对象 *o* 的 *eps* 邻域是以 *o* 为中心、以 *eps* 为半径的空间. 邻域的大小由参数 *eps* 确定, 邻域的密度用邻域内的对象数度量. 通过另一参数 *minPts*, 即指定稠密区域的密度阈值, 来衡量邻域是否稠密. 如果一个对象的 *eps* 邻域至少包含 *minPts* 个对象, 则称该对象为核心对象. 通过连接核心对象及其邻域内其他核心对象的邻域, 形成稠密区域作为簇.

算法1. DBSCAN

输入: 数据集*D*, 半径*eps*, 核心点邻域内点最少数目
输出: 所有簇集合

```
1. 标记D中每一个点为unvisited
2. for each a in D
3. if (a is unvisited)
4.    if (o is not核心点) 标记点o为噪音;
5.    else //a is核心点
6.       点o及其邻域E内的所有点形成一个新簇C;
7.       for (each p in a的邻域)
8.          if (p is unvisited)
9.             if (p is核心点)
10.               将点p的邻域内的所有点加入簇C;
11.            else
12.               if (p没有加入其它任何一个簇)
13.                  将p加入簇C;
14.               if (p被标记为噪音) 取消标记p为噪音;
15. end
```

3.2 基于改进 ReliefF 的无监督特征选择算法 UFS-IR

UFS-IR (Unsupervised Feature Selection based on Improved ReliefF) 算法通过 DBSCAN 聚类算法给样本标注聚类标签来指定样本数据的类别, 为将 ReliefF 算法应用到无监督特征选择提供数据基础. 采用 DB INDEX (Davies Bouldin index, DBI) 准则来判断 DBSCAN 聚类的有效性, 选择 DBI 值较小的一组 (*eps*, *minPts*) 参数对数据进行聚类. 使用不同采样策略控制每个类别样本被抽样总次数 ξm, 确保 m 次抽样中不出现重复抽取样本, 弥补 ReliefF 算法随机选样时小类别样本对更新特征权重的影响容易被忽略这一不足的同时充分发挥每一次采样对特征评估的效用. 至此改进的 ReliefF 算法得到分类相关的特征集合 T 保证最终得到的特征子集 F 的最大相关性, 但是它仍旧无法保证特征子集内部的低冗余度, 在对含有类别信息的数据处理时有些学者使用分类器的正确率来筛选特征[15,16], 但是无法适用无监督特征选择中. 所以本文改进 ReliefF 使用 ACS 系数来衡量特征间的相关性, 结合集合 T 保留强相关特征对中权值较大的特征来保障 F 的最小冗余性.

UFS-IR 算法描述如下:

定义已知包含 n 个特征 $X_1, X_2, \cdots, X_n$ 的数据集 D, 其中 $X_i=(S_1, S_2, S_3, \cdots, S_k)$; 使用聚类算法 DBSCAN, 分别为 D 中 k 个样本划分类别指定标签 Y, 使用改进 ReliefF 评价各个特征 X_i 对于这种分类的影响, 计算 ACS(X_i, X_j) 衡量特征间的相关性, 削减特征子集的冗余度, 获取最终特征子集$F \subsetneq D$.

输入: 样本数据集D, 迭代次数n, 邻域半径*eps*, 样本阈值*minPts*, 抽样次数m, 最近邻样本的个数k, ACS阈值λ
输出: 最优特征子集

1. 初始化所有特征的权值为0, 标记所有样本点为*unvisited*
2. 使用DBSCAN算法处理D, 计算本次聚类的*DBI*
3. for i=2 to n
4. 调整(*eps*, *minPts*), 计算DBI_2
5. if (DBI_2<*DBI*) *DBI*=DBI_2并记录当前(*eps*, *minPts*)
6. 以最终(*eps*, *minPts*)作为DBSCAN的参数对D进行聚类, 标记聚类结果, 删除噪音得到数据集D'
7. 初始化每个聚类簇在m次抽样中被抽中的次数$n=\xi m$, 其中ξ为各聚类簇样本占D'中总样本数的比例
8. for i=1 to m
9. 从D'中随机抽取一个标记为*unvisited*的样本R;
10. 判断R所属类别允许抽中次数的剩余值n是否等于0;
11. if (R所属类别的n==0) 回到(5);
12. else将R标记为*visited*;
13. 从R的同簇样本集中寻找k个最近邻样本$NH_j(j=1, 2, \cdots, k)$, 从R的每一个不同簇样本集中寻找k个最近邻样本$NM(C)_j(j=1,2,\cdots,k)$;
14. for X=1 to N //N为特征总数
15. 根据公式(1)更新特征权值;
16. 删除权值为负数或小于非负权值的均值的特征, 将剩余特征按权值降序排列得到集合T;
17. 基于D, 一个特征的所有值形成一个向量, 计算T中特征两两间的ACS系数;
18. 保留值大于等于阈值λ的特征对*pairs*集合Q;
19. for each *pairs* in Q
20. if (*pairs*中的元素都在T中)
21. 保留权值较大的特征Y;
22. if(F不包含Y) 将Y加入集合F;
23. 输出F, 得到最优特征子集;
24. end

4 实验分析

4.1 实验数据

为了体现算法的有效性, 实验选择了如表 1 所示来自 UCI (University of California Irvin) 机器学习数据库和多伦多大学 Delve Datasets 的四个不同规模数据集, 数据质量较高基本无量纲的影响. 使用支持向量机 SVM 对特征选择后的数据进行分类, 以分类正确率来验证算法的有效性, 使用 10 折交叉验证的结果作为最终分类正确率.

表 1 数据集介绍

数据集	训练集	测试集	属性数	类别数
plice	1000	2175	60	2
madelon	2000	600	500	2
satimage	3104	2000	36	6
mnist	60 000	10 000	780	10

4.2 参数的设置

UFS-IR 算法涉及众多参数, 其中影响算法性能的主要包括 ReliefF 中的抽样次数 m 和 DBSCAN 中的邻域半径 *eps*、样本阈值 *minPts*. 如 2.2 节所述参数 m 将影响最终得出的各个特征的权值, 理论上而言 m 的值增加, 算法将获得更多数据本身的结构信息或统计信息等, 使得特征评估结果更准确. 事实上也是如此, 实验改变对 satimage 数据集的抽样次数 m, 当 m=100 时, UFS-IR-Accuracy 为 0.895; 当 m=200 时, UFS-IR-Accuracy 为 0.922, 当 m=300 时, UFS-IR-Accuracy 为 0.931. 分类正确率随 m 的增加而增加, 但同时运行时间也在增加. 出于实验目的是验证 UFS-IR 对 ReliefF 算法如 2.2 节所述其他三点的改进可行性和有效性, 本

文对于参数 m 以保障最终分类结果优良, 运行时间精炼为原则适当选择.

DBSCAN 对输入参数 (*eps*, *minPts*) 敏感, 若参数选择不当将造成聚类质量下降, 并最终影响对特征评估的准确性. DBSCAN 参数的选择通常依赖于经验, 但当数据内部结构较复杂时, 事先确定合适参数值是比较困难的. 针对这种情况, 在聚类过程中使用 DB INDEX 评价聚类结果, 并以此为依据适当调整参数. 以 satimage 数据集的测试集为例, 设置调整次数 n=10(n 是一个与聚类结果无关的变量, 其值可适当调整) 分析不同 (eps, $minPts$) 参数值相应的 DB INDEX 值以及最终使用 UFS-IR 算法的分类正确率, 如表 2 所示 DB INDEX 对不同参数下聚类效果的评价值越低, 对特征选择的影响更积极使得相应的分类正确率更高, 说明使用 DB INDEX 方法来确定聚类的参数是有效准确的.

表 2 DBSCAN 聚类评估情况

(eps, minPts)	DBI	UFS-IR-Accuracy
(2.35,4)	16.065	0.871
(2.35,5)	18.641	0.828
(2.45,4)	16.894	0.849
(2.45,5)	14.642	0.922
(2.45,6)	15.325	0.891
(2.5,4)	14.795	0.913
(2.5,5)	16.354	0.859

4.3 实验与结果分析

(1) 验证抽样策略和 ACS 有效性

以 splice 数据集为基础, 以原始类别信息作为类别标签, 对比分析 ReliefF 算法和改进抽样策略的 ReliefF 算法对各特征的评估情况如图 1 所示.

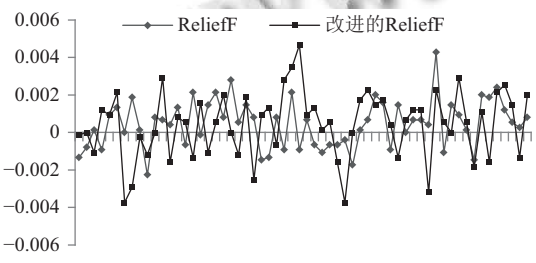

图 1 ReliefF 和改进 ReliefF 对 splice 中 60 个特征的评估情况

图 1 中, 横坐标代表 splice 的 60 个特征, 纵坐标代表两个算法对每个特征的评估值. 从图 1 可以看出, 改进的 ReliefF 对特征的评估各特征间的相对大小和整体趋势与 ReliefF 的评估结果基本一致, 因此改进 ReliefF 算法不会改变特征对于分类的影响因子. 改进抽样策略的 ReliefF 对各特征的评估值明显突出于 ReliefF 的评估结果, 说明不重复抽样充分发挥了每次抽样样本的使用价值, 既定各类样本的抽样概率, 也使得小类别样本稳定发挥作用, 增加了结果准确性和合理性的可靠程度, 证明对 ReliefF 抽样策略的改进是有效的.

ReliefF 家族算法对所有特征进行评估计算相应的权重然后以降序排列, 选择排序后特征的相应比例作为特征子集. 以 splice 和 satimage 数据集的原始规模和类别信息分别训练各自的 SVM 分类模型 model, 对比分析使用 ReliefF, 改进的 ReliefF 以及改进的 ReliefF 结合 ACS 三种方法后, model 对不同维度的数据集的分类正确率.

图 2 和图 3 中纵坐标表示特征维度, 即 ReliefF 体系算法对所有特征排序之后选取前百分之多少构成最终特征选择子集, 横坐标表示分类正确率. 可以看出改进的 ReliefF 相比 ReliefF 算法分类率有所提高, 使用 ACS 删除冗余特征后, 分类率的提高更加明显. 实验说明改进的抽样策略和 ACS 的使用是有效的、可靠的.

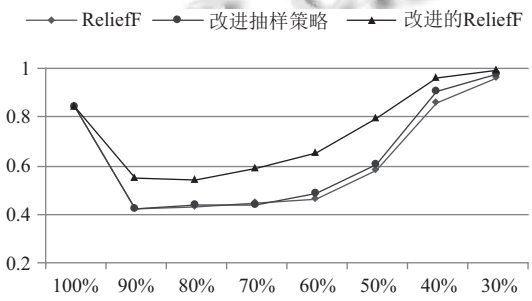

图 2 Splice 数据集各维度分类正确率

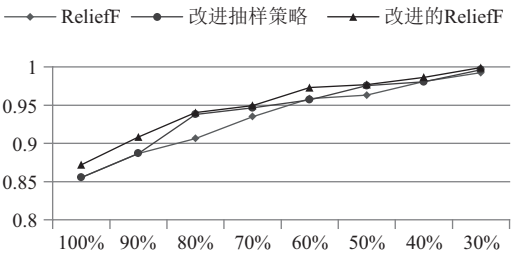

图 3 Satimage 数据集各维度分类正确率

(2) 验证 UFS-IR 的有效性

UFS-IR 算法对原始数据集的特征有两个删减过程, 第 1 次是在完成改进 ReliefF 算法之后, 删除特征评估权重为负值的特征, 负值意味着该特征对于分类具备消极影响. 第 2 次是使用 ACS 对特征间相关性进行衡量, 删减大于阈值 λ 特征对中特征评估权重较小的特征. 本文设定的阈值 λ 为 0.75, 相关性系数不小于 0.75 表示两个特征是显著相关关系. 图 4 展示了各数据集原始特征维度 Original, 第 1 次删减后特征维度 First-cut 以及第 2 次删减后特征维度 Second-cut 的情况. 从图 4 可以看出, UFS-IR 显著删减了原始数据集中的特征数, 使得最终特征子集符合最大相关最小冗余准则.

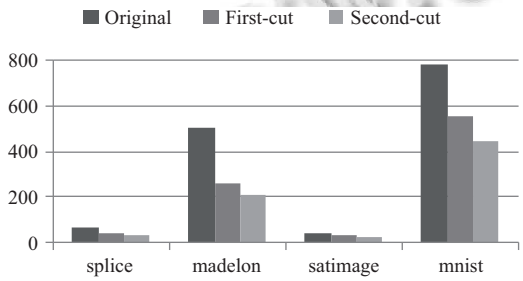

图 4 UFS-IR 特征选择前后维度对比

以各数据集的原始规模和类别信息分别训练各自的 SVM 分类模型 *model*, 并得到对原始数据集的分类正确率 Accuracy. 为验证 UFS-IR 的有效性, 忽略原始数据的类别信息作为 UFS-IR 的实验数据集. DBSCAN 聚类迭代 n=10 次, 选择 DB INDEX 值较小的一组 (*eps*, *minPts*) 参数对数据进行聚类, 并以聚类结果指定数据的类别信息. 用与 *model* 相同的参数对 UFS-IR 特征选择后数据集进行训练和测试, 得到分类正确率 UFS-IR-Accuracy. 如表 3 所示 4 个不同数据集的 UFS-IR- Accuracy 值明显大于 Accuracy, 实验证明本文提出的 UFS-IR 无监督特征选择方法是合理的、有效的.

表 3 特征选择前后分类正确率

数据集	DBSCAN参数 (*eps*, *minPts*)	Relief参数 (*m*, *k*)	ACS阈值λ	Accuracy	UFS-IR-Accuracy
splice	(8.7, 4)	(100, 50)	0.75	0.841	0.9595
madelon	(925, 4)	(200, 50)	0.75	0.835	0.98
satimage	(2.45, 5)	(200, 80)	0.75	0.856	0.922
mnist	(580, 7)	(500, 150)	0.75	0.883	0.9315

5 结论与展望

本文首先分析了无监督特征选择的基础体系, 然后介绍了 ReliefF 算法并对其不足进行分析, 基于此提出改进方案. 本文将 DBSCAN 与改进 ReliefF 加以融合, 提出了基于改进 ReliefF 的无监督特征选择方法 UFS-IR, 构成无监督特征选择的基本结构体系. 实验结果证明, 使用 DB INDEX 准则判定 DBSCAN 有效性并确定参数的方法是有效的, 这也使得 DBSCAN 指导分类更加准确和可靠, 抽样策略的改进使得特征评估的结果更加可信和合理, 调整的余弦相似度有效解决了改进 ReliefF 算法无法辨识冗余特征的不足, 方法间优势互补更提高了 UFS-IR 的有效性和可靠性. UFS-IR 涉及众多参数, 选择更快速有效的方法确定 DBSCAN 的参数, 研究改进 ReliefF 中的参数 *m* 对特征评估的影响以及参数 *m* 的选择策略或准则是今后工作的重点.

参考文献

1 Liu H, Yu L. Toward integrating feature selection algorithms for classification and clustering. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(4): 491–502.

2 Yu L, Liu H. Efficient feature selection via analysis of relevance and redundancy. The Journal of Machine Learning Research, 2004, 5(12): 1205–1224.

3 姚旭, 王晓丹, 张玉玺, 等. 特征选择方法综述. 控制与决策, 2012, 27(2): 161–166, 192.

4 蔡哲元, 余建国, 李先鹏, 等. 基于核空间距离测度的特征选择. 模式识别与人工智能, 2010, 23(2): 235–240.

5 徐峻岭, 周毓明, 陈林, 等. 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.

6 欧璐, 于德介. 基于拉普拉斯分值和模糊 C 均值聚类的滚动轴承故障诊断. 中国机械工程, 2014, 25(10): 1352–1357. [doi: 10.3969/j.issn.1004-132X.2014.10.015]

7 Tabakhi S, Moradi P, Akhlaghian F. An unsupervised feature selection algorithm based on ant colony optimization. Engineering Applications of Artificial Intelligence, 2014, (32): 112–123. [doi: 10.1016/j.engappai.2014.03.007]

8 Yuan HN, Wang SL, Li Y, *et al*. Feature selection with data field. Chinese Journal of Electronics, 2014, 23(4): 661–665.

9 Zhu PF, Zhu WC, Hu QH, *et al*. Subspace clustering guided unsupervised feature selection. Pattern Recognition, 2017, (66): 364–374. [doi: 10.1016/j.patcog.2017.01.016]

10 张丽新, 王家廞, 赵雁南, 等. 基于 Relief 的组合式特征选择. 复旦学报 (自然科学版), 2004, 43(5): 893–898.

11 Kira K, Rendell LA. The feature selection problem: Traditional methods and a new algorithm. Proceedings of 10th National Conference on Artificial Intelligence. San Jose, CA, USA. 1992. 129–134.

12 Kononenko I. Estimating attributes: Analysis and extensions of RELIEF. Proceedings of the European Conference on Machine Learning. Catania, Italy. 1994. 171–182.

13 Sarwar B, Karypis G, Konstan J, *et al*. Item-based collaborative filtering recommendation algorithms. Proceedings of the 10th International Conference on World Wide Web. Hong Kong, China. 2001. 285–295.

14 Jain AK, Murty MN, Flynn PJ. Data clustering: A review. ACM Computing Surveys, 1999, 31(3): 264–323. [doi: 10.1145/331499.331504]

15 谭台哲, 梁应毅, 刘富春. 一种 ReliefF 特征估计方法在无监督流形学习中的应用. 山东大学学报 (工学版), 2010, 40(5): 66–71.

16 刘杰, 金弟, 杜惠君, 等. 一种新的混合特征选择方法 RRK. 吉林大学学报 (工学版), 2009, 39(2): 419–423.