文章编号:2096-1472(2023)11-0048-04
DOI:10.19644/j.cnki.issn2096-1472.2023.011.011

基于冗余性分析的改进 ReliefF 特征选择算法

李丽君[1,4],张海清[1,4],李代伟[1,4],向筱铭[2],于 曦[3]

(1.成都信息工程大学软件工程学院,四川 成都 610225;
2.四川省气象探测数据中心,四川 成都 610072;
3.成都大学斯特灵学院,四川 成都 610106;
4.四川省信息化应用支撑软件工程技术研究中心,四川 成都 610255)

✉ 2432094015@qq.com; zhanghq@cuit.edu.cn; ldwcuit@cuit.edu.cn; micxiang@foxmail.com; yuxi@cdu.edu.cn

摘 要:为了解决 ReliefF 算法随机抽样会抽取到不具代表性的样本且未考虑特征间相关性的问题,提出基于冗余性分析的 ReliefF 特征选择算法。首先改进 ReliefF 的抽样策略,其次将特征权重序列划分为几个子集,分别利用最大信息系数及 Pearson 系数共同衡量特征相关性,设置相应采样比例剔除冗余特征。将改进算法与其他特征选择算法进行对比,结果表明相较于传统 ReliefF,在 LightGBM(Light Gradient Boosting Machine,轻量级梯度提升机器学习)上的分类准确率可提升 0.63%~12.10%,在 SVM(Support Vector Machine,支持向量机)上的分类准确率可提升 0.92%~9.06%,改进算法的分类准确率明显优于其他几种特征选择算法,在考虑特征与标签相关性的同时,能有效剔除冗余信息。

关键词:特征选择;ReliefF 算法;最大信息系数;冗余性分析
中图分类号:TP181 **文献标志码**:A

Improved ReliefF Feature Selection Algorithm Base on Analysis of Redundancy

LI Lijun[1,4], ZHANG Haiqing[1,4], LI Daiwei[1,4], XIANG Xiaoming[2], YU Xi[3]

(1.*School of Software Engineering, Chengdu University of Information Technology, Chengdu 610225, China*;
2.*Sichuan Meteorological Observation and Data Centre, Chengdu 610072, China*;
3.*Stirling College, Chengdu University, Chengdu 610106, China*;
4.*Sichuan Province Engineering Technology Research Center of Support Software of Informatization Application, Chengdu 610225, China*)

✉ 2432094015@qq.com; zhanghq@cuit.edu.cn; ldwcuit@cuit.edu.cn; micxiang@foxmail.com; yuxi@cdu.edu.cn

Abstract: This paper proposes a ReliefF feature selection algorithm based on redundancy analysis to solve the problem of randomly selecting non-representative samples without considering the correlation between features in the ReliefF algorithm. Firstly, the sampling strategy of ReliefF is improved, and then the feature weight sequence is divided into several subsets. The maximum information coefficient and Pearson coefficient are used to jointly measure feature correlation, and corresponding sampling ratios are set to eliminate redundant features. Comparing the improved algorithm with other feature selection algorithms, the results show that compared to traditional ReliefF, the classification accuracy of the improved algorithm can be improved by 0.63% ~12.10% on LightGBM (Light Gradient Boosting Machine), and improved by 0.92% ~9.06% on SVM(Support Vector Machine). The classification accuracy of the improved algorithm is significantly better than other feature selection algorithms, and it can effectively eliminate redundant information while considering the correlation between features and labels.

Key words: feature selection; ReliefF algorithm; maximum information coefficient; analysis of redundancy

0 引言(Introduction)

特征选择是机器学习以及数据挖掘领域实现特征约简的重要方法,通过在众多特征中筛选出对分类最有效的特征实现对特征维数的约简。ReliefF 算法[1]是在 Relief 特征选择算

收稿日期:2023-03-01
基金项目:欧盟项目(598649-EPP-1-2018-1-FR-EPPKA2-CBHE-JP);国家自然科学基金项目(61602064);四川省科技厅项目(2021YFH0107,2022YFS0544,2022NSFSC0571);成都信息工程大学科技创新能力提升计划项目,面向大规模医疗数据的疾病风险评估预测优化研究(KYQN202223)

法[2]的基础上对处理多分类问题提出的改进，但仍存在一些有待解决的问题，例如 ReliefF 随机抽样时会抽取到不具代表性的样本，没有考虑特征间的相关性，缺乏对冗余特征进行衡量。针对以上问题，陈平华等[3]以互信息度量特征冗余。项颂阳等[4]将 ReliefF 与 RFE(Recursive Feature Elimination，特征递归消除)结合对冗余特征进行递归筛选。薛瑞等[5]引入量子粒子群算法对特征集二次筛选剔除冗余特征。张小内等[6]结合 ReliefF 和 Pearson 系数的相关性原理进行特征筛选。此外，已有的对特征间相关性度量的算法评价方式过于单一。

本文提出一种两阶段特征选择算法：①针对样本冗余问题，对 ReliefF 算法抽样策略进行改进，第一阶段保留距各类别中心较近的样本为随机抽样候选集，保证抽取样本的有效性；②针对特征间冗余问题，第二阶段将改进抽样策略后的 ReliefF 算法所得特征权重序列划分为多个区段，在区段内进一步衡量特征间相关性，剔除冗余特征；③引入最大信息系数(Maximal Information Coefficient，MIC)[7]及 Pearson 相关系数共同实现冗余特征的度量；④根据特征权重序列，从高到低给各区段设置采样比例，同时在缩减特征维数的基础上，防止剔除有效特征。

1 ReliefF 算法及其改进(ReliefF algorithm and its improvement)

1.1 Relief 算法

Relief 算法的主要思想是利用特征和类别相关性，根据样本与同类近邻和异类近邻样本的距离相应地更新特征权重。Relief 算法从训练集 $D=\{(x_n,y_n)\}_{n=1}^{N}$ 中随机选取样本 R_i，在 R_i 的同类样本中找出其最近邻样本 NH_i，在 R_i 的异类样本中找出其最近邻样本 NM_i，根据公式(1)更新特征权重：

$$w(j):=w(j)+\frac{diff(j,R_i,NM_i)}{m}-\frac{diff(j,R_i,NH_i)}{m} \tag{1}$$

其中：$w(j)$表示第 j 个特征的权重，m 为随机抽取样本次数，函数 $diff(\cdot)$用于计算在第 j 个特征下两样本点的差值。

1.2 ReliefF 算法

1994 年 Knonenko 提出 Relief 扩展算法 ReliefF[1]，改进后的算法可用于处理多分类问题。ReliefF 公式中针对随机选取的样本是从其同类和异类样本中查找 k 个近邻样本，通过求均值更新特征权重，其公式如下：

$$w(j):=w(j)-\sum_{j=1}^{k}\frac{diff(j,R_i,NH_i)}{mk}+\sum_{c\neq class(R_i)}\frac{\frac{p(c)}{1-p(class(R_i))}\sum_{j=1}^{k}diff(j,R_i,NM_i)}{mk} \tag{2}$$

其中：R_i 为随机抽取的样本；$p(c)$为类 c 的先验概率，即类 c 在样本中所占的比例。

1.3 改进的 ReliefF 算法

1.3.1 冗余样本分析

计算特征权重时，ReliefF 算法需要在整个样本集中进行随机样本的抽取，根据所抽样本与其近邻样本的距离，按照一定规则更新特征权重，随机抽取的样本中存在一些冗余的、不具代表性的样本会一定程度地影响分类结果。

针对上述问题，本文对 ReliefF 随机抽样策略进行改进，在保持抽样随机性不变的前提下，计算各类样本与其类别中心的距离，保留距离所属类别中心较近的部分样本作为随机抽样的候选集，实现对样本抽样范围的缩减，从而避免抽取到一些冗余的、不具代表性的样本，可有效改进 ReliefF 算法衡量特征权重的准确度和最终分类性能。

1.3.2 冗余特征分析

ReliefF 通过特征与标签相关性度量权重，但强相关特征间可能存在冗余[8-9]。故本文引入 MIC 及 Pearson 相关系数分别从信息论[10]和相关性度量[11]两个方面出发共同度量冗余特征。同时，使用两种度量方式避免算法衡量特征相关性时受限于某一度量标准的局限性和盲目性。

MIC 由 RESHEF 等[7]提出，假定存在变量 X、Y，其最大信息系数计算公式如下：

$$MIC(X;Y)=\max_{NM<B}\frac{I(X;Y)}{\log_2\min(M,N)} \tag{3}$$

Pearson 相关性系数主要用于衡量两变量间的相关程度，其中 X、Y 表示两个待测变量，P 为两个变量的相关系数，r 值在 $-1\sim1$，其绝对值越大，表示两个变量间相关性越大，Pearson 系数计算公式如下：

$$P(X,Y)=\frac{\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\overline{Y})^2}} \tag{4}$$

本文对冗余特征的判断使用 MIC 和 Pearson 相关系数共同作为评价指标，将冗余性计算公式定义如下：

$$PM(X,Y)=\alpha\cdot|P(X,Y)|+\beta\cdot MIC(X;Y) \tag{5}$$

假定给定一组特征集 $F=\{f_1,f_2,\cdots,f_m\}$，其中 $\forall f_i\in F$，$i=1,2,\cdots,m$，特征 f_i 的冗余性大小即为特征与子集中其他特征相关性之和，将其定义如下：

$$R(f_i)=\sum_{j=1}^{m}PM(f_i,f_j),j\neq i \tag{6}$$

1.3.3 RFSR 算法

基于上文对样本冗余及特征冗余性的分析，本文在改进样本抽样策略的基础上衡量两两特征之间的相关性，通过将原始特征划分为若干个区段，对不同区段分别剔除冗余特征，提出基于冗余性分析的 ReliefF 算法(ReliefF Feature Selection Algorithm Based on Analysis of Redundancy，RFSR)。

第一阶段，保留距离类中心较近样本作为 ReliefF 算法随机抽样候选集。对于给定一组特征集 $F=\{f_1,f_2,\ldots,f_m\}$，使用 ReliefF 算法获得特征权重序列 $S=\{f'_1,f'_2,\ldots,f'_m\}$。第二阶段，将 S 划分为 h 个区段，假设给定 m 维特征，则每个子集的特征维数为 $m'=\lceil m/h\rceil$，此时第 i 个子集中的特征序列 $S_i=\{f_{i1},f_{i2},\cdots,f_{im'}\}$。在各子集内部使用公式(5)衡量特征间的相关性，相关性越大，说明存在冗余信息的可能性越大，因此按公式(6)计算每个特征冗余大小并升序排序，此时在第 i 个子集中可获得新的特征序列 $S'_i=\{f'_{i1},f'_{i2},\ldots,f'_{im'}\}$。在各子集中根据比例选出部分冗余较小的特征，并组成最终特征集。设第 i 个子集采样比率为 P_i，选取 S'_i 中前$\lceil m/h\rceil*P_i\rceil$个特征，此时第 i 个子集中的特征维数为 $m'=\lceil\lceil m/h\rceil*P_i\rceil$。为确保获

取与标签强相关的特征，权重较高区段获取特征数应多于次要区段，因此采样比率从高到低依次减小设置，例如对于 3 个子集，采样比例可设置为{0.6，0.3，0.1}、{0.5，0.3，0.2}、{0.4，0.4，0.2}。

RFSR 算法的主要思想如下。

(1)计算样本与所属类中心的距离，仅保留距每类中心较近样本作为 ReliefF 随机抽样的候选样本集，缩小随机抽样范围，避免抽取到冗余样本；(2)使用 ReliefF 算法衡量权重，得到特征权重序列；(3)根据所得权重序列将特征进行分段，并从高到低地设置采样比例；(4)在各区段中，使用 Pearson 相关系数及 MIC 组合计算特征间的相关性并升序排序，根据所设采样比率剔除冗余特征，从不同区段获取特征集，保证各子集的多样性。该算法在确保得到更多与标签强相关特征的前提下，剔除出冗余性较高的特征，避免使用单一度量方式时的局限性和盲目性，兼顾特征重要性及冗余性的关系。改进算法伪代码如下。

算法 1：RFSR 算法

输入：训练集 D，取样次数 a，各类样本选取比例 $b\%$，特征个数 m，最近邻数 k，划分区段个数 h，每个区段内特征个数 m'，第 i 个分段的采样比例 P_i，$i=1,2,\cdots,h$，特征权重向量 W。

输出：特征子集 DT。

(1)初始化 $w(i)=0$。

(2)计算各个类别的类中心。

(3)计算每个样本与各自类中心的距离。

(4)按距离由小到大对类别样本进行排序，取各序列中前 $b\%$的样本组成 D'。

(5)FOR $i=1:m$。

(6)FOR $j=1:a$。

(7)在 D'中随机抽取样本 R_i。

(8)找到与 R_i 同类的 k 个最近邻样本 NH_i。

(9)对 $c\neq class(R_i)$，分别找到与 R_i 不同类的 k 个最近邻样本 NM_i。

(10)根据公式(1)更新特征权重 $w(i)$。

(11)END FOR。

(12)END FOR。

(13)根据特征权重排序，得到特征权重序列 S。

(14)将特征序列 S 平均划分为 h 个区段，其中 S_i 表示第 i 个区段。

(15)FOR EACH f_i IN S_i。

(16)根据公式(5)计算特征间的相关性并升序排序，按采样比例 P_i 在每区段特征序列中选取出一组新的特征子集 S_i'。

(17)END FOR EACH。

(18)将各区段中所得特征子集合并形成一组新的特征集 DT。

2 实验结果与分析(Experiment and result analysis)

本文选取 8 个 UCI 公开数据集进行实验对比(表 1)。其中：WDBC 为 Breast Cancer Wisconsin (Diagnostic)数据集，QSAR 为 QSAR biodegradation，Wine 为 Winequality-red，Genus 为 Frogs calls-genus (genus)，Family 为 Frogs calls-family(family)，Heart 为 Statlog(Heart)[12]。

表 1 实验数据集

Tab.1 Experimental dataset

数据集	样本数	特征数	类别数
Sonar	208	60	2
Heart	270	13	2
WDBC	569	32	2
QSAR	1 055	41	2
Wine	1 599	12	7
Spam	4 601	57	2
Genus	7 195	22	4
Family	7 195	22	8

为验证改进算法的有效性，本文进行两组实验，均采用 10 次 10 折交叉验证，将 10 次实验的分类准确率均值作为评价指标，并保留距各类中心较近的前 20%的样本，将冗余性度量公式(5)中的 α、β 值均设为 0.5。实验一中，将不同划分区段、采样比例在不同数据集下进行实验对比，对 10 次实验所得分类准确率求均值，实验一所得结果如表 2 所示。其中：RFSR-6211 和 RFSR-532 分别指划分为 4 个子集和 3 个子集，并将采样比例分别设置为{0.6，0.2，0.1，0.1}和{0.5，0.3，0.2}；加粗数据为最好结果，带下划线数据为第二好结果。

表 2 实验一：不同采样比例下平均准确率对比

Tab.2 Experiment 1:Comparison of average accuracy under different sampling ratios

数据集	分类准确率/%					
	RFSR-6211	**RFSR-4222**	**RFSR-5311**	**RFSR-532**	**RFSR-631**	**RFSR-442**
Sonar	59.90	**67.90**	60.38	63.16	<u>63.17</u>	61.66
Heart	77.75	76.75	77.41	<u>78.09</u>	**78.10**	<u>78.09</u>
WDBC	93.67	93.32	93.49	**94.74**	<u>93.93</u>	93.72
QSAR	77.66	74.91	78.57	<u>79.77</u>	**79.87**	78.83
Wine	51.84	55.84	<u>56.03</u>	51.99	**56.46**	55.97
Genus	91.12	91.08	<u>91.13</u>	**91.15**	**91.15**	91.01
Family	**99.98**	99.95	99.93	<u>99.97</u>	<u>99.97</u>	99.96
Spam	<u>91.08</u>	85.92	88.87	90.26	<u>91.08</u>	**91.84**
平均	80.38	80.71	80.73	81.14	**81.72**	<u>81.39</u>

由表 2 可看出：从区段划分来看，将特征划分为 3 个子集的分类效果整体上要优于 4 个子集；从采样比例来看，采样比例设置为{0.6，0.3，0.1}时，分类效果提升更明显；第一个子集采样占比较高时，所得分类准确率相对较高，还要兼顾后续区段减少特征冗余对分类效果的影响。根据实验一所得结论，实验二将特征序列划分为 3 个子集，采样比例设置为{0.6，0.3，0.1}。将需预设特征个数的对比算法特征数设置为在该比例下所获得的特征数，把 RFSR 与 ReliefF、MIM、mRMR、RF、CFS 以及改进算法 ReliefF-REF[4] 和 ReliefF-Pearson[6] 分别在 SVM 以及 LightGBM 的平均分类准确率进行对比。实验二的实验结果如表 3、表 4 所示。

表3　实验二:不同特征选择算法在 SVM 的分类准确率对比

Tab.3 Experiment 2:Comparison of classification accuracy of different feature selection algorithms under SVM

数据集	分类准确率/%							
	ReliefF	MIM	RF	mRMR	CFS	ReliefF-Pearson	ReliefF-RFE	RFSR
Sonar	59.52	58.73	**66.88**	65.95	58.45	60.12	62.89	63.17
Heart	69.04	77.44	77.43	70.13	69.23	71.63	73.54	**78.10**
WDBC	88.75	78.69	92.71	92.44	88.32	90.32	91.03	**93.93**
QSAR	70.84	67.83	**81.90**	70.28	69.20	73.02	73.32	79.87
Wine	51.84	54.63	53.16	55.53	55.59	52.86	55.92	**56.46**
Genus	89.82	78.10	90.12	89.95	88.05	90.03	90.45	**91.15**
Family	99.05	99.94	99.05	99.05	98.96	99.28	99.83	**99.97**
Spam	85.24	66.85	90.25	84.12	89.62	86.32	89.79	**91.08**

表4　实验二:不同特征选择算法在 LightGBM 的分类准确率对比

Tab.4 Experiment 2:Comparison of classification accuracy of different feature selection algorithms under LightGBM

数据集	分类准确率/%							
	ReliefF	MIM	RF	mRMR	CFS	ReliefF-Pearson	ReliefF-RFE	RFSR
Sonar	58.04	56.72	**71.21**	69.73	56.45	58.43	63.02	64.52
Heart	66.03	**79.85**	77.73	66.28	66.52	72.13	72.21	78.13
WDBC	88.10	84.35	93.85	94.02	89.16	89.20	89.14	**94.37**
QSAR	77.12	74.65	**83.38**	73.26	71.35	77.54	78.50	80.44
Wine	50.84	53.85	54.60	50.28	53.32	52.02	53.29	**55.53**
Genus	90.84	79.23	90.94	90.55	88.05	91.37	90.95	**91.47**
Family	98.94	99.97	98.84	99.37	98.93	98.97	99.49	**99.98**
Spam	83.00	67.61	89.37	84.09	88.95	84.18	88.32	**91.45**

综上可以看出,RFSR 算法在大多情况下的分类准确率优于其他几种特征选择算法,除在 Sonar、QSAR 数据集上 RFSR 算法的分类准确率稍低于 RF 等外,在其他数据集上的分类效果明显更具优势;与经典 ReliefF、mRMR、RF、MIM、CFS 算法相比,RFSR 算法所选特征分类性能更好,并且均高于改进算法 ReliefF-RFE、ReliefF-Pearson;从分类器选择来看,LightGBM 模型分类准确率整体高于 SVM 支持向量机,RFSR 算法使用 LightGBM 在减少特征维度的同时,有效地提高了分类准确率;RFSR 相较于传统 ReliefF 算法,在不同数据集上的分类准确率均有提升,在 SVM 的不同数据集上的分类准确率分别提升 0.92%～9.06%,在 LightGBM 的分类准确率分别提升 0.63%～12.10%,在一定程度上改进了 ReliefF 算法的分类性能。

3　结论(Conclusion)

本文首先对 ReliefF 算法抽样策略进行改进,通过计算类中心缩减随机抽取样本的范围。针对特征间冗余问题,将特征序列划分多个子集,通过两种相关系数共同衡量特征相关性,使 ReliefF 同时兼顾特征与标签及特征间的关系,消除冗余特征的不良影响。在 8 个 UCI 数据集上展开实验对比,通过实验确定参数设置,同时分别在 SVM 及 LightGBM 上将改进算法与其他几种算法进行对比。结果表明:改进算法在降低特征维度的同时,能有效提高分类准确率,但算法没考虑不平衡数据及算法稳定性问题,若不同类别样本数量差异较大,则可能会影响算法性能。未来,会从不平衡数据性质出发,进一步对算法性能提升展开研究。

参考文献(References)

[1] KONONENKO I. Estimating attributes:analysis and extensions of RELIEF[C]//Springer. Machine Learning:ECML-94. Berlin,Heidelberg:Springer,1994:171-182.

[2] KIRA K,RENDELL L A. The feature selection problem:traditional methods and a new algorithm[C]//ACM. Proceedings of the tenth national conference on Artificial intelligence. New York:ACM,1992:129-134.

[3] 陈平华,黄辉,麦淼,等. 结合 ReliefF 和互信息的多标签特征选择算法[J]. 广东工业大学学报,2018,35(5):20-25,50.

[4] 项颂阳,许章华,张艺伟,等. 高光谱图像分类的 ReliefF-RFE 特征选择算法构建与应用[J]. 光谱学与光谱分析,2022,42(10):3283-3290.

[5] 薛瑞,赵荣珍. ReliefF 与 QPSO 结合的故障特征选择算法[J]. 振动与冲击,2020,39(11):171-176,208.

[6] 张小内,程文鹏,侯惠让,等. 基于 ReliefF-Pearson 的嗅觉脑电通道选择[J]. 电子与信息学报,2021,43(7):2032-2037.

[7] RESHEF D N,RESHEF Y A,FINUCANE H K,et al. Detecting novel associations in large data sets[J]. Science,2011,334(6062):1518-1524.

[8] 施启军,潘峰,龙福海,等. 特征选择方法研究综述[J]. 微电子学与计算机,2022,39(3):1-8.

[9] 刘强,降爱莲. 基于交互信息的两阶段特征选择算法[J]. 计算机工程与设计,2023,44(1):125-132.

[10] 程双勤,刘倩,朱懿敏. 基于最大信息系数的随机森林算法[J]. 信息技术与信息化,2021(7):37-40,46.

[11] 王俊红,赵彬佳. 基于不平衡数据的特征选择算法研究[J]. 计算机工程,2021,47(11):100-107.

[12] MANSOURI K,RINGSTED T,BALLABIO D,et al. Quantitative structure-activity relationship models for ready biodegradability of chemicals[J]. Journal of Chemical Information and Modeling,2013,53(4):867-878.

作者简介:

李丽君(1997-),女,硕士生。研究领域:机器学习。

张海清(1986-),女,博士,教授。研究领域:数据挖掘,机器学习,人工智能。

李代伟(1976-),男,博士,副教授。研究领域:数据挖掘,机器学习和粗糙集。本文通信作者。

向筱铭(1985-),男,硕士,高级工程师。研究领域:气象信息系统研发。

于　曦(1973-),男,博士,教授。研究领域:决策系统,深度学习。