doi:10.3969/j.issn.1001-4616.2021.01.016

基于变异系数和最大特征树的特征选择方法

徐海峰,张　雁,刘　江,吕丹桔

(西南林业大学大数据与智能工程学院,云南 昆明 650224)

[摘要]　特征选择是数据挖掘的关键过程,特征贡献度评分和特征优选是其核心部分. 针对特征贡献度评分,提出一种用变异系数度量类内距离、互信息度量类间距离的 CVMI(coefficient of variation and mutual of information)方法,将该算法运用到嵌入式特征选择方法中进行特征优选. 实验采用 UCI 提供的 4 组数据集、1 组遥感数据和 1 组鸟鸣声数据,使用 7 种特征贡献度评分方法进行对比. 结果表明,CVMI 方法更符合特征贡献度评价的客观规律,对比其他 7 种方法,CVMI 方法取得较好效果. 此外,基于 CVMI 特征评分方法构建最大特征树,结合二邻域去冗余的特征优选方法 CVMI-RRMFT(remove redundancy of maximum feature tree),采用上述数据集进行实验,结果表明该方法不仅能有效降低数据维度,而且还能提高分类准确率.

[关键词]　特征选择,特征贡献度,变异系数,互信息,最大特征树,二邻域去冗余
[中图分类号]TP3-0　[文献标志码]A　[文章编号]1001-4616(2021)01-0111-08

Feature Selection Method Based on Coefficient of Variation and Maximum Feature Tree

Xu Haifeng,Zhang Yan,Liu Jiang,Lü Danjv

(School of Big Data and Intelligent Engineering,Southwest Forestry University,Kunming 650224,China)

Abstract:Feature selection is a key process in data mining. Feature contribution scoring and feature optimization are its core parts. This paper proposed a CVMI(coefficient of variation and mutual of information)method that used the coefficient of variation to measure the distance between intraclass and the mutual information to measure the distance between interclass,and then applied the algorithm to the embedded feature selection method. The experiment used four UCI data sets,one set of remote sensing data and birds sound data,and tested seven different feature contribution scoring methods. The results showed that the CVMI method was more in line with the objective law of feature contribution evaluation. It also achieved better results compared to the other feature scoring methods. Besides,this paper also proposed a feature optimization method CVMI-RRMFT(remove redundancy of maximum feature tree)based on CVMI to construct a maximum feature tree and remove redundancy with two-neighborhood. Experiment results demonstrated that this feature optimization method effectively reduced data dimensions and improved the classification accuracy.

Key words:feature selection,feature contribution scoring,coefficient of variation,mutual information,maximum feature tree,remove redundancy with two-neighborhood

特征选择一直以来是机器学习和模式识别等领域的热点研究问题[1-2]. 它是数据预处理的必需过程,其目的是在特征空间中挑选出冗余度低、判别性高的特征子集[3],从而避免“维数灾难”的发生,它将直接影响数据分析的时间成本、模型的泛化性和解决目标问题的效果及性能[4].

特征选择可分为两部分:特征贡献度评分和特征优选. 特征贡献度评分是特征选择的关键过程,为了度量数据特征之间的相似性和差异性,按照指定度量规则计算特征对分类空间的贡献度. 从方法而言,可分为 4 类:基于相似度、基于信息论、基于统计方法和基于稀疏学习[5]. 目前,已经有大量研究提出许多不同的特征评分算法[6],包括在无监督特征选择中利用 Laplacian 算子评估特征贡献度[7]、在环境音识别中

收稿日期:2020-09-16.
基金项目:国家自然科学基金资助项目(61462078,31860332)、云南省教育厅科学研究基金资助性项目(2017ZZX212).
通讯作者:张雁,博士,教授,研究方向:智能信息处理、机器学习. E-mail:zydyr@163.com

使用 Constraint Score 进行特征评分[8]、在利用 SVM 诊断心脏病中使用 Fisher Score 筛选特征[9]和基于变异系数对遥感数据进行降维[10]等. 特征优选则是从特征评分序列中挑选出冗余低、易分类的特征子集,该过程可分为封装式、嵌入式和过滤式 3 大类. 近年来不少研究基于这 3 种方式提出了很多优化算法,如使用二进制蜻蜓优化的特征选择[11]、基于粒子群算法的特征优选[12]、基于人工蜂群和梯度提升决策树的特征选择[13]和基于人工蚁群的特征选择[14]等. 文献[11]的方法是将连续特征搜索空间映射到离散空间,并提出 S 变形和 V 变形作为蜻蜓优化算法的传递函数进行特征优选;文献[12]将互信息作为特征贡献度评价指标结合粒子群算法剔除冗余特征;文献[13]使用文献[14]中提出的梯度增强决策树作为特征贡献度评价的方法,再利用人工蜂群对评价后的特征子集进行特征优选. 文献[15]以 Pearson 相关系数对特征评分,使用人工蚁群对特征进行优选.

目前,计算特征贡献度的方法都是基于空间距离度量特征之间的相关性和差异性,如 Constraint Score 和 Laplacian,虽然算法复杂度低,但忽略了特征之间量纲不一致的问题,而且由于数据样本类内离散度不一致,随机抽样对结果也会产生影响;另外一些算法从概率统计学角度,用互信息对特征进行评分,例如文献[12]计算每一个特征样本与分类标签的互信息,虽然解决了量纲不同的问题,但是忽略了类内差异. 此外,分类标签为离散数据,特征样本有离散数据也有连续的,未经处理直接计算某特征和标签列的互信息是不合理的. 近年来,许多研究将特征优选过程看作优化过程[6],并结合许多智能算法搜索目标特征,但是对于数据量较大的数据集,多目标优化算法时间、空间复杂度比较高,而且利用多目标优化算法降维增加了模型的敏感性,降低其泛化性.

针对上述问题,本文从最小化类内距离和最大化类间距离的角度出发,利用变异系数和互信息的特点,提出 CVMI(coefficient of variation and mutual of information)特征贡献度评分方法,将其融入嵌入式特征优选中,不仅能解决量纲不同的问题,而且能有效减少特征,提高建模精度;此外,本文还提出了一种基于 CVMI 特征评分结合最大生成树和二邻域去冗余的特征优选 CVMI-RRMFT(remove redundancy of maximum feature tree)方法,该方法不仅能剔除冗余特征,有效降低算法的复杂度,而且能提高模型的精度和泛化性.

1　CVMI 特征评价

在分类问题中,分类结果的好坏取决于选定特征的可分性. 特征优选原则:类内距离小、类间差异大. 本文提出一种联合类内变异系数 CVAC(coefficient of variation for intraclass)和类间互信息 MIIC(mutual of information for interclass),构建联合指标 CVMI(coefficient of variation and mutual of information)来评价特征贡献度,结合嵌入式方法进行特征选择.

基于 CVMI 的特征选择过程如图 1 所示,步骤如下:

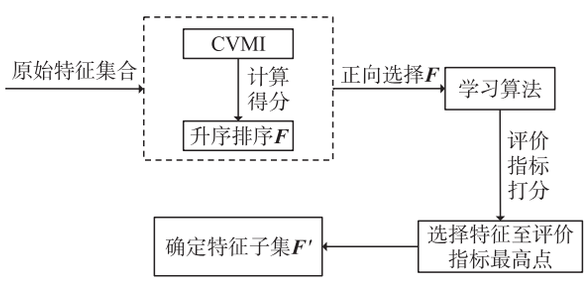

图 1　CVMI 特征选择方法

Fig. 1　The method of CVMI feature selection

Step 1:对原始特征使用 CVMI 方法对其评分,升序排序记作 $\boldsymbol{F}$,初始化特征子集 $\boldsymbol{F}'=\varnothing$;

Step 2:从 $\boldsymbol{F}$ 中正向选择一个特征,纳入 $\boldsymbol{F}'$中;

Step 3:将 $\boldsymbol{F}'$映射至训练集 $\boldsymbol{T}$r(Train Dataset)和测试集 $\boldsymbol{T}$e(Test Dataset),得映射集合 $\boldsymbol{T}$r′、$\boldsymbol{T}$e′;

Step 4:使用分类器对 $\boldsymbol{T}$r′、$\boldsymbol{T}$e′建模和分类评价,将评价指标 k(准确率或 Kappa 等)记录到 $\boldsymbol{K}$ 中;

Step 5:重复 Step 2 至 Step 4 过程,直到 $\boldsymbol{F}'$包含 $\boldsymbol{F}$ 中所有元素;

Step 6:找到 $\boldsymbol{K}$ 中最大值对应的特征个数 m,将 $\boldsymbol{F}$ 中的前 m 个特征作为最终特征子集.

1.1　类内变异系数 CVAC

在统计学中,变异系数是用来度量两个或多个观测值样本的变异程度,也可以用于测量它们之间的离散度. 其表达式如下:

$$Cv=\sigma/\mu. \tag{1}$$

式中,μ、σ 分别为样本的均值和方差. 在特征空间为 $\boldsymbol{F}$,分类空间为 $\boldsymbol{C}$ 的问题中,特征 $f(f\in\boldsymbol{F})$ 的类内变异

系数为:

$$CVAC_f=\sum_{i=1}^{C}Cv_i. \tag{2}$$

式中,Cv_i 为第 f 个特征中第 i 类样本的变异系数,$CVAC_f$ 表示 $\boldsymbol{C}$ 中第 f 个特征的类内变异系数. 当 $CVAC_f$ 越大,则第 f 个特征的离散度越高,反之,则第 f 个特征较为内聚.

1.2 类间互信息 MIIC

在信息论中,互信息用于度量两个变量的依赖或相关程度. 对于两个随机离散变量 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$,互信息 $I(\boldsymbol{X};\boldsymbol{Y})$ 的计算表达式为:

$$I(\boldsymbol{X};\boldsymbol{Y})=H(\boldsymbol{X})-H(\boldsymbol{X}|\boldsymbol{Y})=\sum_{y\in\boldsymbol{Y}}\sum_{x\in\boldsymbol{X}}p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right). \tag{3}$$

式中,$p(x,y)$ 表示 x 和 y 的联合概率分布函数,$p(x)$ 和 $p(y)$ 表示 x 和 y 的边缘概率密度. $I(\boldsymbol{X};\boldsymbol{Y})$ 的取值区间为[0,1],$I(\boldsymbol{X};\boldsymbol{Y})$ 等于 0 时说明两个变量相互独立;$I(\boldsymbol{X};\boldsymbol{Y})$ 越大,则 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 联系越紧密.

通常而言,用互信息评价属性贡献度时,$\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 分别为某特征向量和标签向量[12,16],但是特征向量和标签向量数据类型存在不一致:特征向量多为连续数据,而标签向量多为离散数据. 本文用互信息度量类间距离,$\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 表示为同一特征下不同类的样本向量. 若分类空间和特征空间同 $CVAC$,则第 f 个特征的类间互信息 $MIIC_f$ 如下:

$$MIIC_f=\sum_{i\in\boldsymbol{C}}\sum_{j\in\{\boldsymbol{C}\setminus i\}}I(i;j). \tag{4}$$

式中,i, $j(i\neq j)$ 分别为第 f 个特征中第 i 类和第 j 类的样本,$MIIC_f$ 为 $\boldsymbol{F}$ 中第 f 个特征的类间互信息. $MIIC_f$ 越小,则第 f 个特征类间差异越大,反之亦然.

1.3 协同指标 CVMI

根据属性评价原则:类间差异大,类内距离小. 本文提出 CVMI 用于评价特征贡献度,计算表达式为:

$$CVMI_f=\lambda CVAC_f+(1-\lambda)MIIC_f. \tag{5}$$

式中,对于特征优选,$CVAC_f$、$MIIC_f$ 都要小. 考虑到度量类内和类间指标权重不一致,引入调节参数 λ,平衡二者的权重.

如图 2,在分类空间 $C=\{A,B,C\}$ 的样本空间中,A 类样本的特征 $\boldsymbol{X}$ 与 $\boldsymbol{Y}$ 的类内距离为 $CVAC_{\boldsymbol{X}}$、$CVAC_{\boldsymbol{Y}}$. 显然,$CVAC_{\boldsymbol{Y}}$ 相对 $CVAC_{\boldsymbol{X}}$ 较离散;在特征 $\boldsymbol{X}$ 中,B 类样本和 C 类样本的类间距离为 $MIIC(B;C)_{\boldsymbol{X}}$,即 B 类和 C 类映射在特征 $\boldsymbol{X}$ 上的互信息;如果随机变量相互独立,则互信息为 0,如图 2 中 A 类样本与 B 类样本在特征 $\boldsymbol{X}$ 上的映射交集为∅. 由此可见特征 $\boldsymbol{X}$ 下样本可分性较高.

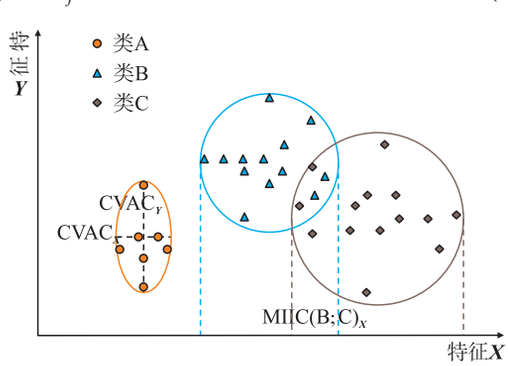

图 2 CVMI 示意图

Fig. 2 The schematic of CVMI

2 CVMI-RRMFT

除对特征进行贡献度评分外,还要考虑特征之间的相关性. 本文基于特征评分与去冗余提出 CVMI-RRMFT(remove redundancy of maximum feature tree)的方法,利用 CVMI 贡献度评分序列构建最大特征树,再根据 RRMFT 方法剔除冗余特征.

该方法属于过滤式特征选择,流程如图 3 所示. 计算原始特征集相关系数邻接矩阵,构建最大特征树 $\boldsymbol{T}$;同时对每个特征使用 CVMI 算法得到贡献度评分序列 $\boldsymbol{F}$. 联合 $\boldsymbol{T}$、$\boldsymbol{F}$ 根据二邻域方法剔除冗余特征,确定特征子集.

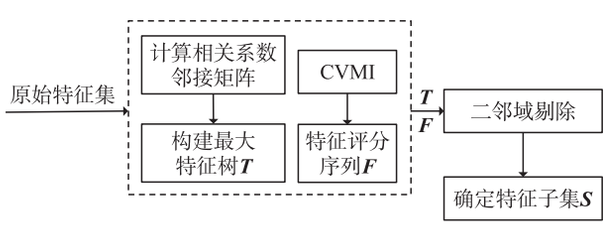

图 3 CVMI-RRMFT 流程图

Fig. 3 The flow diagram of CVMI-RRMFT

2.1 构建最大特征树

最大特征树,根据最小生成树衍生而来. 在

无向图 $\boldsymbol{G}(\boldsymbol{V},\boldsymbol{E})$ 中,每边都有一个权值 w,最小生成树则是在 $\boldsymbol{G}$ 中选取一个边集合 $\boldsymbol{E}'$,使得在 $\boldsymbol{V}$ 中所有点都被 $\boldsymbol{E}'$ 所连接,同时让 $\boldsymbol{E}'$ 中所有边权值之和最小. 最大特征树中的节点 $\boldsymbol{V}$ 为特征属性,使得 $\boldsymbol{E}'$ 中所有边权值之和最大,边权值由 Pearson 相关系数决定. $\boldsymbol{P}_{(\boldsymbol{F}_r,\boldsymbol{F}_c)}$ 表示特征 $\boldsymbol{F}r$ 和 $\boldsymbol{F}c$ 的相关性,其计算表达式:

$$\boldsymbol{P}_{(\boldsymbol{F}_r,\boldsymbol{F}_c)}=\frac{\sum_i(\boldsymbol{F}_{ri}-\boldsymbol{F}_r)(\boldsymbol{F}_{ci}-\boldsymbol{F}_c)}{\sqrt{\sum_i(\boldsymbol{F}_{ri}-\boldsymbol{F}_c)^2}\sqrt{\sum_i(\boldsymbol{F}_{ci}-\boldsymbol{F}_c)^2}}. \tag{6}$$

$$\boldsymbol{I}_{(\boldsymbol{F}_r,\boldsymbol{F}_c)}=\frac{-\log_2(1-\boldsymbol{P}^2_{(\boldsymbol{F}_r,\boldsymbol{F}_c)})}{2}. \tag{7}$$

式(6)中 $\boldsymbol{F}_{ri}$ 表示第 r 个特征的第 i 个样本点,$\boldsymbol{F}_r$ 表示第 r 个特征样本的均值. 式(7)中 $\boldsymbol{I}_{(\boldsymbol{F}_r,\boldsymbol{F}_c)}$ 表示属性 $\boldsymbol{F}_r$、$\boldsymbol{F}_c$ 的带权相关度. 由式(6)、(7)计算出相关系数邻接矩阵,构建最大特征树,算法 MaxFT(Maximum Feature Tree)描述如算法 1.

算法 1　构建最大特征树 MaxFT

算法名称:构建最大特征树 MaxFT

输入:相关度矩阵 $\boldsymbol{D}_{n\times n}$;($n$ 为特征个数)

过程:(1)初始化无向图 $\boldsymbol{G}=\{1\}$

(2)while $|\boldsymbol{G}|\leqslant n$ do

(3)　$c=n\backslash\boldsymbol{G}$

(4)　$m=\boldsymbol{D}$ 的$\{\boldsymbol{G}\times\{c\}\}$子矩阵中最大值列标

(5)　$\boldsymbol{G}=\boldsymbol{G}\cup m$

(6)end while

(7)return $\boldsymbol{G}$

输出:最大特征树 $\boldsymbol{G}$

2.2　二邻域剔除冗余

根据最大特征树,结合 CVMI 特征贡献度评分方法计算得出评分序列(按照评分方法指定规则进行排序),剔除冗余特征,操作步骤如图 4 所示.

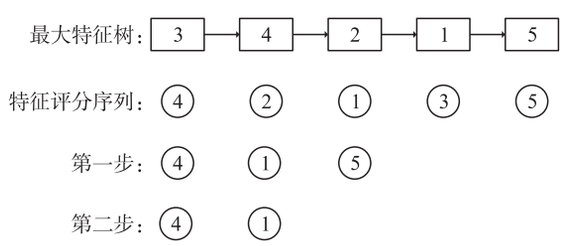

图 4　二邻域剔除冗余特征

Fig. 4　Remove the redundant features with two-neighborhood

示例:如图 4,CVMI 特征评分序列 $\boldsymbol{F}=\{f_4,f_2,f_1,f_3,f_5\}$,遍历 $\boldsymbol{F}$,如第一个特征 $x=f_4$.

Step 1:在最大特征图中寻找 x 所在位置,得出相邻元素 $m=\{f_3,f_2\}$,在 $\boldsymbol{F}$ 中将 m 删除,得出 $\boldsymbol{F}'=\{f_4,f_1,f_5\}$;

Step 2:依次遍历 $\boldsymbol{F}'$中未遍历的元素,取 $x=1$,同理可得 $m=\{f_2,f_5\}$,在 $\boldsymbol{F}'$中将 m 删除,$\boldsymbol{F}'=\{f_4,f_1\}$;

Step 3:$\boldsymbol{F}'=\{f_4,f_1\}$ 该集合元素已经全部遍历,则最终特征序列为$\{f_4,f_1\}$.

3　实验

3.1　实验准备

(1)实验数据

本实验选用 6 组数据集,其中 4 组数据集为 UCI[17] 提供:Sonar,Vehicle,Diabetes 和 Ionosphere;另外 2 组数据为遥感数据 RS 和鸟鸣数据 Birds. 每组数据按 7∶3 的比例划分训练集和测试集. 数据集信息如表 1 所示.

表 1　数据集信息

Table 1　Data set information

数据集名称	训练样本数	测试样本数	特征个数	类别总数
Sonar	156	52	60	2
Vehicle	634	212	18	4
Diabetes	576	192	8	2
Ionosphere	263	88	33	2
RS	1 239	633	25	5
Birds	10 426	4 466	39	6

(2)实验设置

在实验中,为验证 CVMI 和 CVMI-RRMFT 方法的有效性,将它们与 Constraint 得分(CS)[8] 和 Weka 平台提供的 6 种特征选择方法:Correlation(Cor)、GainRatio(GR)、InfoGain(IG)、OneR(OR)、ReliefF(RF)、

SymmetricalUncert(SU)作对比. 并设置 CVMI 计算式中的 λ 值为 0.8.

(3)分类器与性能评价

本实验以决策树为分类器,对特征贡献度评分序列 $\boldsymbol{F}$(每种评分方法按照指定规则排序,如 CVMI 为升序)正向选择,映射训练集和测试集,评价指标使用 Kappa 系数,其表达为:

$$Kappa=\frac{p_o-p_e}{1-p_e}. \tag{8}$$

式中,p_o 为总体分类精度(Acc),即所有正确样本个数 m 除于样本总数 n,如式(9):

$$p_o=m/n. \tag{9}$$

p_e 基于混淆矩阵计算而来:假定每一类真实样本个数为 $a_1,a_2,\cdots,a_n$,每一类预测样本为 $b_1,b_2,\cdots,b_n$,则有

$$p_e=\frac{a_1\times b_1+a_2\times b_2+\cdots+a_n\times b_n}{n\times n}. \tag{10}$$

在本实验中每组数据独立重复运行 10 次,并计算 10 次的 Kappa 系数的均值.

3.2 CVMI 实验结果与分析

在图 5 中,CVMI 方法为带 * 号标记的线条. 从 Kappa 峰值可以看出 CVMI 方法能有效提高建模精度:CVMI 方法 Kappa 峰值相比其他 7 种方法可达最高,在 Sonar、Vehicle 和 Diabetes 数据上,分别提高 2.65%、3.10%和 1%;在 Ionosphere 和 RS 数据集上,CVMI 较其他方法而言使用特征数最少就能达到精度峰值,图 5(f)效果更为显著,由该图可看出本文 CVMI 方法均比其他 7 种选取很少特征就能达到 Kappa 峰值,符合特征优选和贡献度排序准则;其次,特征优选:计算特征贡献度排序后,根据该序列正向选择特征并进行分类测试,分类精度原则上应先是上升趋势,在达到峰值后,由于纳入的特征对模型可能存在一定干扰,则开始呈现波动趋势. 在图 5 中从 Kappa 曲线趋势来看:将 Kappa 峰值作为截止项,CVMI 较上升趋势较为稳定,而其他 7 种方法波动幅度大.

在表 2 中,CVMI 方法除在 Diabetes 数据集上性能不明显,在其他数据集上均是使用最少的特征数 Kappa 值就能达到最高值,尤其在 60 维的 Sonar 数据和 33 维的 Ionosphere 数据上效果显著. 在 Sonar 数据集上,7 种对比方法中 SymmetricalUncert 使用特征数最少为 28 个特征,而其他 6 种方法均为 50 个特征以上,但 SymmetricalUncert 对应 Kappa 值为 0.81,而 CVMI 仅用 21 个特征 Kappa 值就达到 0.87,对比

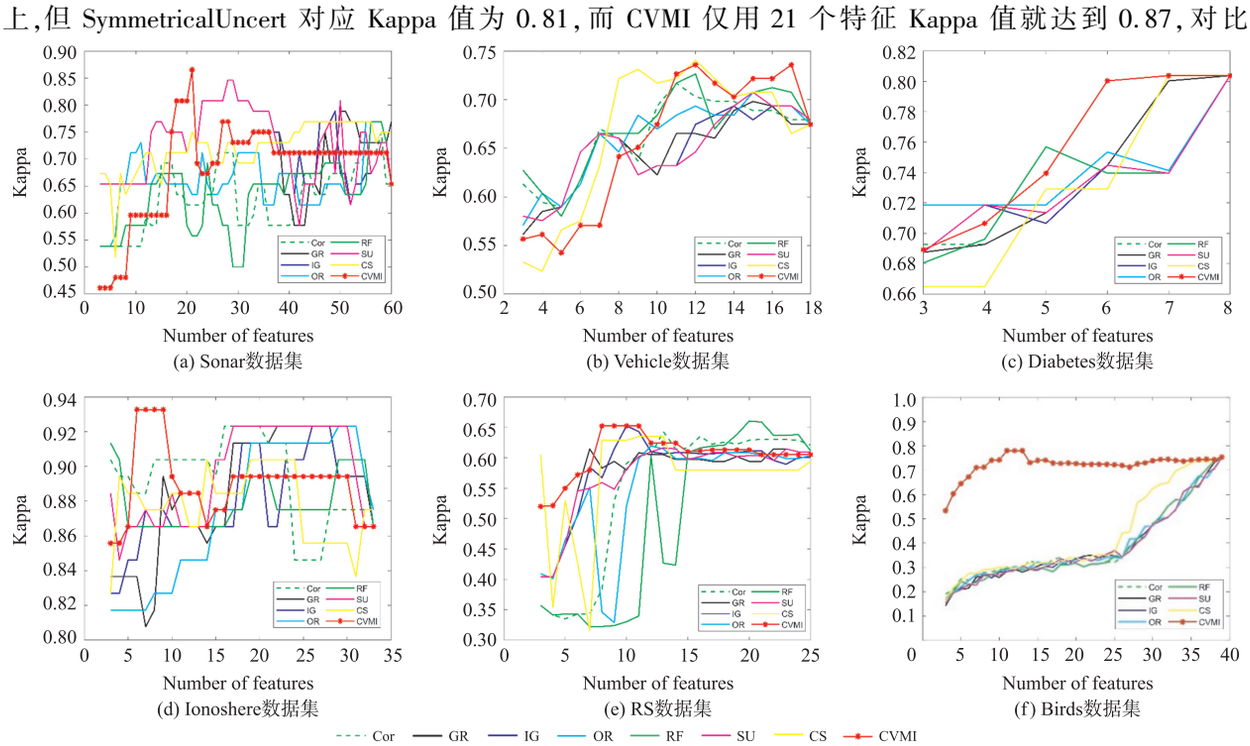

图 5 CVMI 正向特征选择评估

Fig. 5 CVMI method forward feature selection evaluation

SymmetricalUncert 方法增长了 6%;在遥感数据集 RS 上 GainRatio 方法使用特征数最少,使用 7 个特征计算 Kappa 值为 0.63,对于 CVMI 虽然用了 8 个特征,但 Kappa 值对比 GainRatio 方法提高了 3%;表现效果最佳的是 Birds 数据集,仅用了 9 个特征达到了 Kappa 峰值,对比其他 7 种方法,降维率为 75%. 对于 RS 数据集 ReliefF 方法精度和本文方法相同,但 ReliefF 使用了 20 个特征,CVMI 仅使用 8 个特征. 同时还对原始数据进行实验对比,可以看出,CVMI 不仅能有效地降低维度,而且还有效提高了建模精度.

表 2　不同数据集使用不同特征贡献度评价方法对比结果

Table 2　Comparison results of different data sets with different feature contribution evaluation methods

方法	数据集名称					
	Sonar	Vehicle	Diabetes	Ionosphere	RS	Birds
	达到 Kappa 峰值特征数(个)|Kappa 峰值					
Correlation	58|0.75	11|0.71	8|0.80	24|0.92	13|0.64	37|0.76
GainRatio	56|0.76	15|0.70	8|0.80	15|0.92	7|0.63	36|0.75
InfoGain	50|0.78	14|0.69	8|0.80	10|0.92	10|0.65	37|0.76
OneR	55|0.77	15|0.71	8|0.80	28|0.92	12|0.62	35|0.76
ReliefF	56|0.77	15|0.71	8|0.78	22|0.92	20|0.66	34|0.75
SymmetricalUncert	28|0.81	15|0.71	8|0.80	21|0.92	13|0.62	37|0.76
ConstraintSorce	56|0.77	13|0.72	7|0.79	20|0.92	12|0.62	39|0.76
CVMI	21|0.87	12|0.74	7|0.80	6|0.93	8|0.66	9|0.78
原始数据	60|0.74	18|0.68	8|0.78	33|0.92	25|0.61	39|0.76

3.3　CVMI-RRMFT 实验结果与分析

在 CVMI-RRMFT 选择特征方法实验中,使用 3.1 中 Weka 6 种特征评价方法:Cor、GR、IG、OR、RF、SU 和 CS[8] 方法以及本文方法 CVMI,计算特征贡献度评分序列,将该序列结合去冗余方法 RRMFT;分类器采用决策树 J48,对选定的特征子集在训练集和测试集上映射. 每组实验独立重复 10 次,并计算 10 次准确率的均值.

实验结果如图 6,横坐标表示 7 种特征评价方法和本文 CVMI 方法,ORI 为原始数据;纵坐标表示 10 次准确率的均值;柱状图正上方对应的数字为选定的特征个数. 图 6 中,RRMFT(采用 8 种不同特征贡献度评分)方法对比原始数据,6 组数据都表明 RRMFT 不仅能有效地降低维度,而且有效提高了分类准确率.

在表 3 中,对 6 组数据而言,8 种特征评分方法结合 RRMFT 都将维度降低至 50%到 55%,在 Sonar、Vehicle、RS 和 Birds 数据集上,都是使用最少的特征数达到最高的准确率;本文 CVMI-RRMFT 方法 Sonar 的降维率为 55%,准确率较原始数据提高 12%,Birds 降维率为 53.85%,准确率较原始数据提高 9.71%;在 Ionosphere 数据集上准确率 InfoGain 最高,本文 CVMI 方法与其相差 1.13%,但降维率是最高的. 综上所述,CVMI-RRMFT 在降维和提高模型准确率上效果显著.

表 3　不同数据集使用不同特征贡献度评价方法

Table 3　Different feature contribution evaluation methods with different data sets

方法	数据集名称					
	Sonar	Vehicle	Diabetes	Ionosphere	RS	Birds
	去冗余后特征数(个)|准确率/%					
Correlation-RRMFT	27|82.62	10|71.26	3|74.19	18|93.18	13|64.40	23|81.34
GainRatio-RRMFT	27|75.00	9|70.75	3|74.19	19|94.32	11|67.14	23|80.00
InfoGain-RRMFT	28|75.00	10|70.75	3|74.19	19|95.45	11|62.87	21|79.69
OneR-RRMFT	28|80.79	9|64.15	5|72.67	18|94.32	10|63.66	29|82.50
ReliefF-RRMFT	27|73.07	9|70.75	5|72.67	18|92.05	13|60.82	23|86.64
SymmetricalUncert-RRMFT	27|75.00	10|70.50	3|74.19	19|92.05	11|62.87	22|85.40
CS-RRMFT	28|73.07	9|70.75	3|69.76	19|90.91	13|60.98	21|83.69
CVMI-RRMFT	27|85.00	8|71.26	5|76.04	12|94.32	10|67.14	18|86.40
ORI	60|73.08	18|67.92	8|71.51	33|94.32	25|60.98	39|76.69

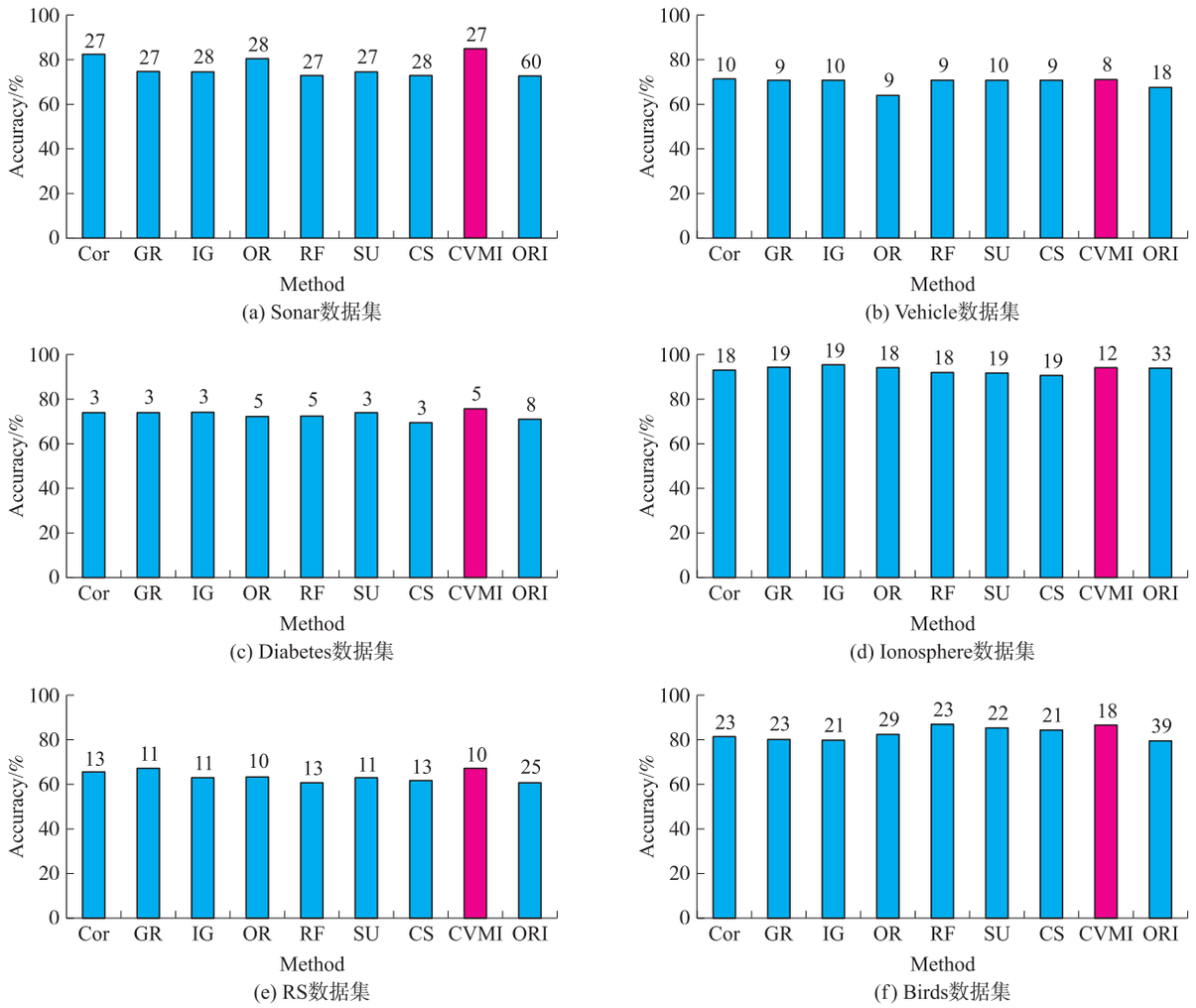

图 6　不同数据集的 RRMFT 特征选择性能

Fig. 6　Accuracy of RRMFT feature selection with different data sets

此外,本文还对 CVMI 和 CVMI-RRMFT 两种特征选择方法进行了实验对比. 在实验中,使用 CVMI 和 CVMI-RRMFT 方法进行特征选择后,将 Kappa 和准确率(Acc)作为评价指标. 如表 4 所示,6 组数据集 CVMI-RRMFT 方法准确率都高于 CVMI 方法,其中 Sonar 数据集的效果最为显著;Acc 高出 4.35%,Kappa 高出 0.04;在数据 RS 中 CVMI 和 CVMI-RRMFT 的表现一致;在 Vehicle 数据集中 Acc 指标相同,CVMI 的 Kappa 系数对比 CVMI-RRMFT 仅高出 0.01. 综上所述,CVMI-RRMFT 方法对比单独使用 CVMI 性能更强.

表 4　不同数据集使用 CVMI 和 CVMI-RRMFT 方法的实验结果

Table 4　Experimental results of CVMI and CVMI-RRMFT methods used in different data sets

方法	评价指标	数据集名称					
		Sonar	Vehicle	Diabetes	Ionosphere	RS	Birds
CVMI	Kappa	0.79	0.71	0.72	0.87	0.63	0.83
	Acc(%)	80.65	71.26	72.67	93.18	67.14	85.40
CVMI-RRMFT	Kappa	0.83	0.70	0.76	0.90	0.64	0.80
	Acc(%)	85.00	71.26	76.04	94.32	67.14	86.40

4　结论

大数据时代,降维是不可或缺的步骤. 对此,许多研究将特征评价和去冗余作为多目标,从而提出不同的优化算法,但是对于海量数据而言,时间成本是非常高的. 另外,在特征相关性分析时数据存在不同量纲. 本文提出了嵌入式 CVMI 特征选择方法,使用变异系数和互信息度量类内和类间距离,解决了属性

量纲不同的问题,能有效选择较优的特征提高分类精度,而且对比其他特征评选方法,更符合特征评选的客观趋势;此外,将 CVMI 与去冗余方法 RRMFT 结合,综合考虑特征贡献度和特征相关性,通过实验证明该方法能有效遴选出可分性高、冗余度低的特征. 实验使用了 UCI 提供的 4 组公共数据集和两组自建数据,本文方法在这些数据集上表现都是良好的,更好地结合特征评价和冗余去除也是未来的研究方向.

[参考文献]

[1] Kozodoi N,Lessmann S,Papakonstantinou K,et al. A multi-objective approach for profit-driven feature selection in credit scoring[J]. Decision support systems,2019,120:106-117.

[2] JIANG B,LI C,RIJKE M D,et al. Probabilistic feature selection and classification vector machine[J]. ACM transactions on knowledge discovery from data,2019,13(2):1-27.

[3] KULKARNI A,METTA R. A new code obfuscation scheme for software protection[C]//2014 IEEE 8th International Symposium on Service Oriented System Engineering. Oxford:IEEE,2014:409-414.

[4] COLLBERG C,THOMBORSON C,LOW D. A taxonomy of obfuscating transformations[D]. New Zealand:The University of Auckland,1997.

[5] LI J,CHENG K,WANG S,et al. Feature selection:a data perspective[J]. ACM computing surveys,2017,50(6):1-45.

[6] 李郅琴,杜建强,聂斌. 特征选择方法综述[J]. 计算机工程与应用,2019,55(24):10-19.

[7] ZHANG Y,WANG Q,GONG D,et al. Nonnegative Laplacian embedding guided subspace learning for unsupervised feature selection[J]. Pattern recognition,2019,93:337-352.

[8] ZHAO S,ZHANG Y,XU H,et al. Ensemble classification based on feature selection for environmental sound recognition[J]. Mathematical problems in engineering,2019,2019.

[9] SAQLAIN S M,SHER M,SHAH F A,et al. Fisher score and Matthews correlation coefficient-based feature subset selection for heart disease diagnosis using support vector machines[J]. Knowledge and information systems,2019,58(1):139-167.

[10] 张康,黑保琴,周壮,等. 变异系数降维的 CNN 高光谱遥感图像分类[J]. 遥感学报,2018,22(1):87-96.

[11] MAFARJA M,ALJARAH I,HEIDARI A A,et al. Binary dragonfly optimization for feature selection using time-varying transfer functions[J]. Knowledge-based systems,2018,161:185-204.

[12] 王金杰,李炜. 混合互信息和粒子群算法的多目标特征选择方法[J]. 计算机科学与探索,2020,14(1):83-95.

[13] RAO H,SHI X,RODRIGUE A K,et al. Feature selection based on artificial bee colony and gradient boosting decision tree[J]. Applied soft computing,2019,74:634-642.

[14] WANG H,MENG Y,YIN P,et al. A model-driven method for quality reviews detection:an ensemble model of feature selection[C]//Wuhan International Conference on E-Business. Wuhan,China,2016:2.

[15] 巫红霞,谢强. 基于加权社区检测与增强人工蚁群算法的高维数据特征选择[J]. 计算机应用与软件,2019,36(9):285-292,301.

[16] 程玉胜,宋帆,王一宾,等. 基于专家特征的条件互信息多标记特征选择算法[J]. 计算机应用,2020,40(2):503-509.

[17] DUA D,GRAFF C. UCI Machine Learning Repository[http://archive.ics.uci.edu/ml]. Irvine,CA:University of California,School of Information and Computer Science. 2019.

[责任编辑:陆炳新]