计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2018.20170748
55(10):2321-2330,2018

一种小样本数据的特征选择方法

许 行[1] 张 凯[1] 王文剑[1,2]
[1](山西大学计算机与信息技术学院 太原 030006)
[2](计算智能与中文信息处理教育部重点实验室(山西大学) 太原 030006)
(xuh102@126.com)

A Feature Selection Method for Small Samples

Xu Hang[1], Zhang Kai[1], and Wang Wenjian[1,2]
[1](*School of Computer and Information Technology, Shanxi University, Taiyuan* 030006)
[2](*Key Laboratory of Computational Intelligence and Chinese Information Processing* (*Shanxi University*), *Ministry of Education*, *Taiyuan* 030006)

Abstract For small samples, the common machine learning algorithms may not obtain good results as the feature dimension of small samples is often larger than the number of samples and some irrelevant or redundant features are often existed. It is an effective way to solve this problem by reducing the feature dimension through feature selection. This paper proposes a filter feature selection method based on mutual information for the small samples. First, the criterion of feature grouping based on the mutual information is defined. Both the correlations between features and the class and the redundancy among different features are considered in this criterion, according to which the features are grouped. Then those features that have maximal correlation with the class in each group will be chosen to compose a candidate feature subset. Meanwhile, it is ensured that the time complexity of this algorithm is low. After that, the feature selection method based on feature grouping is combined with Boruta algorithm to determine the optimal feature subset automatically from the candidate feature subset. In this way, the feature dimension can be reduced greatly. Compared with the five classical feature selection algorithms, experimental results on benchmark data sets demonstrate that the feature subset selected by the proposed method has better classification performance and running efficiency on three kinds of classifiers.

Key words small samples; feature selection; mutual information; feature grouping; filter algorithm

摘 要 小样本数据由于其特征维数相对于样本数目较多,且常包含不相关或冗余特征,使得常用的机器学习算法处理小样本数据时无法得到好的效果,通过特征选择来降低数据维数是解决该问题的一种有效途径.针对小样本数据,提出一种基于互信息的过滤型特征选择方法,首先定义了基于互信息的特征分组标准,该标准同时考虑特征与类别的相关性和不同特征之间的冗余性,根据该标准对特征分组

收稿日期:2017-09-30;**修回日期**:2018-03-03
基金项目:国家自然科学基金项目(61673249);山西省回国留学人员科研基金项目(2016-004);赛尔网络下一代互联网技术创新项目(NGII20170601)
This work was supported by the National Natural Science Foundation of China (61673249), the Scientific Research Foundation for Returned Scholars of Shanxi Province (2016-004), and the CERNET Innovation Project (NGII20170601).
通信作者:王文剑(wjwang@sxu.edu.cn)

后,在各组内选出与类别相关性最大的特征构成候选特征子集,保证了算法具有较低的时间复杂度,之后采用 Boruta 算法,在候选特征子集中自动确定最佳特征子集,从而大幅度降低数据的维数.通过与 5 种经典的特征选择算法比较,在标准数据集上采用 3 种分类器的实验结果表明提出的方法选出的特征子集具有较好的运行效率和分类性能.

关键词 小样本数据;特征选择;互信息;特征分组;过滤型算法

中图法分类号 TP181

随着通信和存储技术的发展、网络的普及,各领域数据的产生和收集变的更加容易,大数据及相关产业应运而生,而处理这些数据成为机器学习与数据挖掘领域研究的核心及关键问题.现实生活中,有一类称为小样本的数据,其特点是相比于特征维数其样本数目较少,比如基因表达谱数据需要通过微阵列实验获取,实验成本的昂贵限制了实验次数,使得数据的规模较小,同时该实验测试上万个基因的表达水平,又使得数据维数非常高,这使得传统的机器学习算法处理小样本数据可能会失效[1],因此,通过特征选择来降低数据维数是解决它的一种有效途径.

特征选择能在不失去数据原有价值的基础上去除不相关和冗余特征,提高数据的质量,降低学习算法在数据集上的计算代价,加快数据挖掘的速度,同时有助于生成更易理解的结果和更紧凑、泛化能力更强的模型[2].根据是否与后续完成数据分析任务(如分类、聚类、回归等)的算法相独立,特征选择方法可分为嵌入、封装和过滤 3 类[3].嵌入型方法将特征选择算法作为分类算法的一个组成部分嵌入到分类算法中,封装型方法将后续分类算法的分类准确率作为所选特征子集的评价准则,过滤型方法与后续分类算法无关,直接利用训练数据的统计性能评估特征.对于嵌入型和封装型方法,将特征选择算法作为分类算法的组成部分或者使用分类算法作为特征子集的评价标准,都会造成特征选择算法的计算成本随着维数的升高而急剧上升,可能不适合小样本数据的特征选择.而过滤型方法有独立的评估函数,通过样本的统计属性来评价特征子集对于分类任务所起的作用,它不将任何分类器纳入到评估标准,由此选择出无关于特定分类算法的特征子集.因此,过滤型方法可以离线进行特征选择,它相对于后续分类算法的独立性可避免高维数据造成的较高的分类算法运行成本,与嵌入型和封装型相比,过滤型特征选择方法在计算上是高效的.

典型的过滤型特征选择方法使用距离度量、信息度量、相关性度量和一致性度量等统计指标衡量特征的类区分能力.距离度量是利用距离来度量特征之间、特征与类别之间的相关性,常用的有欧氏距离、S 阶闵可夫斯基测度、切比雪夫距离、平方距离等,Relief[4]及其变种 ReliefF[5]、BFF(best first strategy for feature selection)[6]和基于核空间距离方法[7]都是基于距离度量的算法.信息度量是指选择具有最小不确定性的特征,常用的信息度量为衡量信息不确定性的熵函数,如 Shannon 熵、条件熵、信息增益、互信息(mutual information, MI)等.BIF(best individual features)[8], UFS-MI(unsupervised feature selection approach based on mutual information)[9], CMIM(conditional mutual information maximization)[10]分别是使用互信息和条件互信息作为评价标准的特征选择方法.相关性度量是利用特征与类别的可分离性间的重要性程度判断相关性,如 Pearson 相关系数、概率误差、Fisher 分数、线性可判定分析、最小平方回归误差[11]、平方关联系数[12]等.Ding 等人[13]和 Peng 等人[14]在 mRMR(minimal-redundancy-maximal- relevance)中处理连续特征时,分别使用 F-Statistic 和 Pearson 相关系数度量特征与类别和已选特征间的相关性程度,Hall[15]给出一种同时考虑特征的类区分能力和特征间冗余性的相关性度量标准.一致性度量是指给定 2 个样本,若他们特征值相同而类别不同,则它们是不一致的,否则是一致的,一致性准则试图保留原始特征的辨识能力,用不一致率来度量,典型算法有 Focus[16], LVF(Las Vegas filter)[17]等.这些方法有的运行效率不够高,有的降维之后分类模型性能不够好,因此研究针对小样本数据的过滤型特征选择方法仍有重要的价值.

由于互信息有 2 个优点[18]:1)可以测量随机变量之间的多种关系,包括非线性关系,这保证了互信息在特征与类别之间的关系未知的情况下仍然有效;2)在平移、旋转和保留特征矢量顺序的特征空间变换情况下,值不会发生改变,这保证了互信息在

特征选择中的任何阶段都能准确度量任意 2 个特征之间的关系.因此,基于互信息的过滤型特征选择方法可以很好地度量特征与特征之间、特征与类别之间的关系,从而更有效地进行特征选择.本文针对小样本数据提出一种基于互信息的过滤型特征选择方法,用以提高其选出的特征子集所构造的分类模型的分类性能,同时具有更好的运行效率.

1　小样本数据的特征选择方法

本文首先提出一种基于互信息的特征选择方法(MI-based feature selection, MIFS),根据互信息对特征排序,之后按顺序迭代地对特征分组,在各组内选出与类别相关性最大的特征得到特征子集,然后利用 Boruta 算法[19]自动地确定最佳特征子集.

1.1　基于互信息的特征选择方法

有效的特征选择方法需要同时考虑特征与类别的相关性和不同特征之间的冗余性,并且避免在类别相关度差别较大的特征上计算冗余度.为了实现以上 2 点,提出基于互信息的特征选择算法 MIFS.考虑到互信息度量特征与类别之间的关系的优势,MIFS 先根据特征与类别之间的相关性对特征排序,之后提出了一个分组标准,将特征进行分组,并从不同的组内找到需要选出的特征作为特征子集.

给定数据集 D 的样本数为 n,特征维数为 m,用 $a_1,a_2,\cdots,a_m$ 表示其特征,c 表示其类别,特征 a_i 的值域为 V_i,c 的值域为 V_c.

特征 a_i 与类别 c 之间的互信息 $I(a_i,c)$ 为

$$I(a_i,c)=\sum_{v_i\in V_i}\sum_{v_c\in V_c}p(v_i,v_c)\ \mathrm{lb}\ \frac{p(v_i,v_c)}{p(v_i)p(v_c)},\quad(1)$$

其中 $p(v_i,v_c)$ 表示特征 a_i 的取值为 v_i 且类别 c 的取值为 v_c 的概率.$I(a_i,c)$ 的值越大,表示特征 a_i 和类别 c 的关联度越大.

计算每个特征与类别之间的互信息后,按互信息从大到小的顺序对特征排序,然后对特征集进行分组,定义特征分组的标准 Q 为

$$Q=\frac{S_G}{R_G},\quad(2)$$

$$S_G=\sum_{a_i\in G}I(a_i,c),\quad(3)$$

$$R_G=\sum_{a_i,a_j\in G}I(a_i,a_j),\quad(4)$$

其中,G 表示一个特征组,a_i,a_j 为 G 内的特征,$I(a_i,a_j)$ 为特征 a_i 与特征 a_j 之间的互信息:

$$I(a_i,a_j)=\sum_{v_i\in V_i}\sum_{v_j\in V_j}p(v_i,v_j)\ \mathrm{lb}\ \frac{p(v_i,v_j)}{p(v_i)p(v_j)},\quad(5)$$

其中 $p(v_i,v_j)$ 表示特征 a_i 的取值为 v_i 且特征 a_j 的取值为 v_j 的概率.$I(a_i,a_j)$ 的值越大,表示特征 a_i 和 a_j 越相似.

这里 S_G 为特征组 G 与类别的关联度,R_G 为特征组 G 内所有特征的相似性,特征组 G 的 Q 值越大,表示该特征组中的特征与类别的关联度越大,特征组内特征之间的冗余度越小;反之,Q 值越小,表示该特征组中的特征与类别的关联度越小,特征组内特征之间的冗余度越大.

为了计算特征分组的初始 Q 值,需要选出 2 个特征放入分组中:首先将排在第 1 位的特征 a_1 放入分组,然后计算特征 a_1 与其他每个特征 a_i 之间的互信息,并选出互信息最大的特征,即最相似的特征放入该分组.之后按式(2)计算分组的 Q 值,记录为 q_0.

对于其他特征,将此时排在最前面的特征添加到当前分组中,再计算其 Q 值,如果 $Q<q_0$,说明该特征与当前分组中已有的特征冗余度较高,则继续将下一个特征添加到当前分组,并计算新的 Q 值,迭代直到该特征组的 $Q>q_0$ 时停止向这个特征组添加特征,此时的特征组就作为第 1 个分组.在未被分组的特征上重复上述步骤得到新的特征组,依此类推,直到所有的特征都被分入特征组中,则得到所有特征组.最后取出每个特征组中的第 1 个特征作为其所在特征组的代表,用取出的特征构成候选特征子集.

MIFS 算法的主要步骤总结如下:

算法 1. MIFS 算法.

输入:数据集 D、候选特征个数 k;

输出:候选特征子集 S_{can}.

① 按式(1)计算数据集 D 中每个特征与类别 c 的互信息 $I(a_i,c)$;

② 将特征按互信息从大到小排序,得到特征集 A;

③ 按以下步骤对特征集 A 分组:

④ 令 $t=1$,从 A 中取出排在第 1 位的特征 a_1 放入分组 G_t;

⑤ 按式(5)计算特征 a_1 与其他每个特征 a_i 之间的互信息 $I(a_i,a_j)$,将最大的特征放入分组 G_t;

⑥ 按式(2)计算 G_t 的 Q 值,记为 q_0;

⑦ 从A中剩余的特征中取出排在最前面的特征放入分组G_t中，按式(2)计算G_t的Q值；如果$Q \leqslant q_0$，则重复步骤⑦；如果$Q > q_0$，则将当前的G_t作为第1个分组；

⑧ 令$t=t+1$，在剩下的A上重复步骤④～⑦，得到新的特征组G_t，直到$t=k$，或者A中所有特征都被分入特征组中时停止；

⑨ 取出每个特征组的第1个特征放入特征集S_{can}；

⑩ 返回S_{can}.

1.2 MIFS-Boruta 算法

MIFS算法可以通过对特征分组的方式去除冗余特征，但它同大多数过滤型特征算法类似，无法自动确定最佳特征.

Boruta[19]是一种全相关的封装型特征选择方法，它试图找到携带可用于预测的信息的所有特征，而不是像大多数传统封装型算法一样只找到在分类器上产生最小误差的特征子集. 无论特征与决策变量的相关性强弱与否，Boruta都会找到所有的相关特征，这使得它非常适合应用于确定最佳特征子集.

Boruta算法首先将数据集扩充，通过随机打乱原数据集各特征的取值，生成与原数据集的特征数量相同的“影子”特征，由于这些“影子”特征是随机生成的，所以Boruta算法认为它们是不重要的特征. 之后分别在各个原始特征与“影子”特征上采用随机森林进行分类，计算各特征的效果，将“影子”特征中分类效果最好的特征作为衡量原始特征是否重要的标准，从而去除不重要的特征. Boruta算法能找到候选特征中与类别相关的所有特征，从而直接确定特征数目，得到最优特征子集.

Boruta可以找到所有相关特征这一优点正好可以解决MIFS算法无法自动给出最佳子集的问题，因此我们考虑建立MIFS和Boruta的混合模型. 在混合模型中，封装型算法可以充分利用过滤方法获得的结果，提高运行效率，并获得产生较高分类性能的子集，同时，过滤型方法也可以利用封装型方法来确定特征子集中的特征个数，这样封装和过滤方法的特性得到了很好的互补[14]. 因此本节提出一种基于MIFS和Boruta的混合模型，用以设计高效的特征选择算法自动选出一组冗余较小且数量较小的特征，称为MIFS-Boruta算法.

MIFS-Boruta算法的主要步骤归纳如下：

算法 2. MIFS-Boruta特征选择算法.

输入：数据集D、候选特征子集个数k、迭代次数r；

输出：特征子集S.

① 在数据集D上运行MIFS算法得到包含k个候选特征的特征集S_{can}；

② 从数据集D中取出特征集S_{can}对应的数据作为新的数据集D_{sub}；

③ 在数据集D_{sub}上运行Boruta算法，Boruta算法的迭代次数为参数r；

④ 返回特征子集S.

MIFS算法初始化时选择的和类别互信息最大的特征将会包含在最优的特征子集中，因为该特征首先被放入第1个特征分组，之后根据分组标准Q向该特征组中添加特征使其内部的特征之间有较高的冗余度，同时使得特征组与类别的关联度随着特征数量的增加而减小，所以在此特征组中只需选择一个与类别关联度最大的特征作为该组的代表，这个特征就是算法初始化时选出的和类别互信息最大的特征，因此该特征被选为最优特征子集的候选特征；然后采用Boruta算法从候选特征中去除不重要的特征，而由于与类别互信息最大的特征的分类效果通常不会低于“影子”特征的分类效果，因此该特征不会被去除，包含在最优特征子集中.

1.3 时间复杂度分析

假设给定数据集的样本数为n，特征维数为m，则MIFS算法中求类别和每个特征之间的互信息的时间复杂度为$O(mn^2)$，对特征排序的时间复杂度为$O(m\log m)$，迭代地对特征分组的复杂度在最坏的情况下为$O(mn)$，所以MIFS算法的时间复杂度为$O(m\log m+mn^2)$. 由于本文算法针对小样本数据，其中$n \leqslant m$，因此可以将样本数n视为常数，得到关于特征维数m的时间复杂度为$O(m\log m)$.

而MIFS-Boruta特征选择算法的运行时间是由MIFS算法和Boruta算法2部分运行时间组成，如果用k表示第1阶段MIFS算法得到的候选特征子集中特征的个数，根据文献[19]中的分析可知，Boruta算法的时间复杂度为$O(kn)$，同理，在小样本问题中可看作关于维数的时间复杂度$O(k)$. 综上，MIFS-Boruta特征选择算法的时间复杂度为$O(m\log m)+O(k)$，又因为$k \leqslant m$，因此算法的时间复杂度实际上为$O(m\log m)$.

2　实验结果与分析

2.1　实验数据

为了验证算法在高维数据上的性能和有效性,以及该方法是否适用于实际问题,本文使用了 11 个公开可用的数据集,特征数目在 1 024～19 993 之间,平均特征个数为 6 924,其中 6 个数据集的维度超过了 5 000,3 个数据集具有不少于 10 000 维的特征,这些数据集主要是图像和生物微阵列数据,数据集的详细信息如表 1 所示. 为了便于处理,本文对连续型特征的数据使用等距离散化的方法进行了预处理. 实验在 1 台 i7-2600 3.40 GHz 4 核处理器、4 GB 内存的电脑上运行,开发环境为 Matlab R2015a.

Table 1　Datasets Used in the Experiments

表 1　实验数据集

Data ID	Dataset	#Instances	#Features	#Classes	Data Type
1	ALLAML	72	7 129	2	Biological
2	Carcinom	174	9 182	11	Biological
3	GLIOMA	50	4 434	4	Biological
4	lung	203	3 312	5	Biological
5	orlraws10P	100	10 304	10	Image
6	pixraw10P	100	10 000	10	Image
7	Prostate_GE	102	5 966	2	Biological
8	SMK_CAN_187	187	19 993	2	Biological
9	warpAR10P	130	2 400	10	Image
10	warpPIE10P	210	2 420	10	Image
11	Yale	165	1 024	15	Image

2.2　特征选择结果比较

为了验证本文算法是否能够获得较好的特征选择结果,将本文的 MIFS-Boruta 算法与 CMIM[10], ICAP (interaction capping)[20], CIFE (conditional infomax feature extraction)[21], mRMR[14], L1MI (L1 least-squares mutual information)[22] 5 种经典的特征选择算法进行比较,其中 CMIM, ICAP, CIFE, L1MI 方法是基于互信息度量的过滤型特征选择算法,mRMR 是基于相关性度量的过滤型特征选择算法. 这些方法在使用时一般都要指定降维之后的特征数,为公平起见,实验中将这些方法分别与 Boruta 方法结合,预先设定了特征选择算法在每个样本集上的候选特征数 k,本文根据经验将其设为原始数据集特征维数的 1.5%～5%之间. 表 2 为在

Table 2　Feature Selection Results Under the Combinations of Different Feature Selection Algorithms and Boruta Algorithm

表 2　不同特征选择算法与 Boruta 算法组合时特征选择结果

Dataset	m	k	CMIM	ICAP	CIFE	mRMR	L1MI	MIFS
ALLAML	7 129	150	23	16	2	56	63	50
Carcinom	9 182	200	178	165	29	187	175	137
GLIOMA	4 434	120	23	6	7	33	31	26
lung	3 312	120	120	109	18	120	110	93
orlraws10P	10 304	200	173	168	164	170	172	172
pixraw10P	10 000	200	196	92	93	200	186	200
Prostate_GE	5 966	150	20	27	8	22	30	28
SMK_CAN_187	19 993	300	18	22	9	20	14	26
warpAR10P	2 400	120	74	65	64	68	116	70
warpPIE10P	2 420	120	110	114	114	120	120	120
Yale	1 024	50	31	50	21	35	28	27

不同数据集上各特征选择算法选出的特征个数. 由于每种方法的第 2 阶段都是 Boruta,故本文表中的方法名称都省略了-Boruta.

从表 2 可以看出所有的特征选择算法所选出的特征个数远小于原始特征维度 m,最终选出的特征个数也明显小于候选特征个数,CIFE 算法在 8 个数据集上都得到了最少的特征个数,ICAP 和 CMIM 算法分别在 2 个和 1 个数据集上取得了最少的特征个数,本文的 MIFS 方法所选出的特征个数在 5 个数据集上少于 L1MI 和 mRMR,4 个数据集上少于 CIMI.

5 种算法分别和 Boruta 算法组合的特征选择方法得到的 5 个特征子集中,存在部分与 MIFS-Boruta 算法所选特征相同的特征,相同特征的个数如图 1 所示:

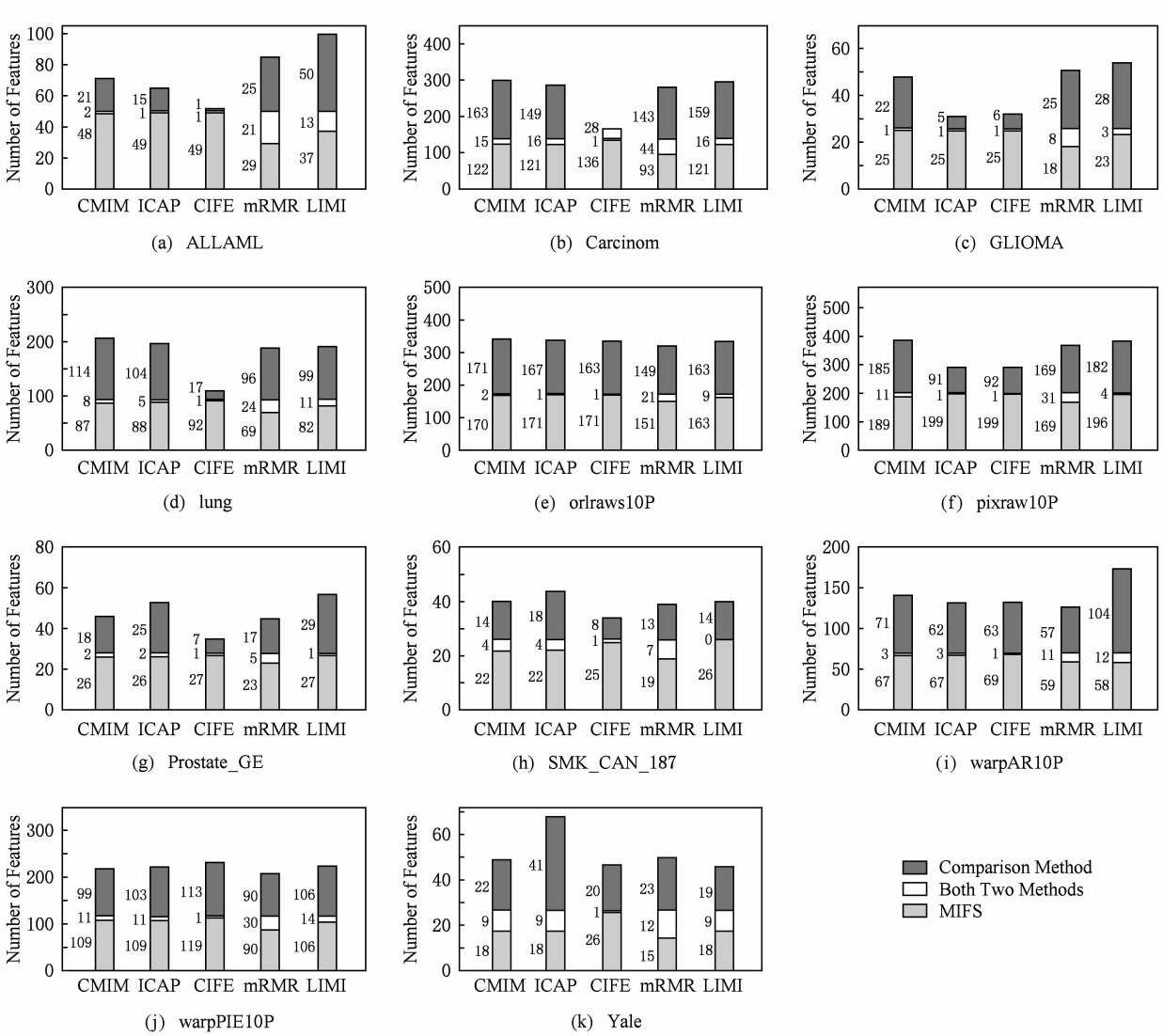


Fig. 1 The comparison of the same features selected by MIFS-Boruta algorithm and other algorithms

图 1 MIFS-Boruta 算法与其他算法选出的相同特征比较

从图 1 可以看出,MIFS 选出的特征与 5 种算法选出的特征中基本上都存在相同的特征,其中与 mRMR 方法所选特征最为相近,相同特征的数量最多. 尽管 MIFS 与这些方法选出的特征不尽相同,但后边的实验表明这对分类结果的影响不大.

2.3 运行时间比较

由于确定最佳特征子集都采用算法 Boruta,因此只需要比较 6 种算法在确定候选特征子集过程中的运行时间,这些算法在 11 个数据集上的运行时间结果如表 3 所示. 为了比较的公平性,本实验将选出的特征数目 k 全部设定为 150 个. 从表 3 中可以看到,所提出的 MIFS 算法在 8 个数据集上的运行时间均小于其他几个基于互信息的特征选择算法,只有在 lung,pixraw10P,SMK_CAN_187 这 3 个数据集上,MIFS 算法稍慢于 CMIM 算法. 因此在大多数数据集上,MIFS 算法具有更高的运行效率.

Table 3　The Running Time of Different Feature Selection Algorithms

表 3　不同特征选择算法运行时间　　s

Dataset	MIFS	CMIM	ICAP	CIFE	mRMR	L1MI
ALLAML	**0.27**	0.28	19.39	19.07	6.31	166.55
Carcinom	**0.42**	0.49	50.79	50.74	12.09	44.90
GLIOMA	**0.20**	0.27	11.66	11.26	3.54	24.27
lung	0.22	**0.07**	2.73	2.64	0.72	67.35
orlraws10P	**0.43**	0.54	42.23	41.77	11.09	76.21
pixraw10P	0.44	**0.42**	35.55	34.91	9.58	71.07
Prostate_GE	**0.25**	**0.25**	18.90	18.31	5.55	956.11
SMK_CAN_187	0.80	**0.42**	92.41	92.64	27.16	166.55
warpAR10P	**0.30**	0.60	11.59	11.37	2.83	24.21
warpPIE10P	**0.12**	0.63	15.01	15.18	3.49	50.89
Yale	**0.40**	0.50	12.59	11.57	3.13	34.49
Average Time	**0.35**	0.41	28.44	28.13	7.77	152.96

2.4　在特征子集的分类性能比较

为了验证所提算法特征选择的有效性,分别选取支持向量机(support vector machine, SVM),决策树,*K*-近邻(*K*-nearest neighbor, KNN) 3 个分类器作为分类算法. SVM 是监督学习模型,本文使用常用的线性 SVM 模型;决策树是通过学习算法构造的树形结构的分类器,它是一种非线性分类器,本文使用经典的 ID3 算法;对于 KNN 分类器,选用 1NN 算法(单最近邻算法),它通过最邻近的 1 个样本的类别来决定待分样本所属的类别. 在所有数据集上进行 10 次十折交叉验证测试分类性能.

实验使用 3 个指标来评价特征子集选择算法的性能:1)最低分类错误率;2)平均最低分类错误率;3)Win/Tie/Lose 记录(该记录表示在给定度量上,所提算法获得比其他特征选择算法更好、相等和更差的性能的数据集数目,可简记为 W/T/L).

不同特征选择方法得到的特征子集在使用 SVM、决策树和 KNN 作为分类器时的最低分类错误率分别如表 4～6 所示,表中的 Average Error 表示各方法在所有数据集下的平均最低分类错误率,W/T/L 行表示所提出方法在 11 个数据集上的分类错误率胜于、相同、弱于其所在列的方法的数据集数目.

Table 4　Classification Error Rate on SVM Classifier

表 4　SVM 分类器上的分类错误率　　%

Dataset	CMIM	ICAP	CIFE	mRMR	L1MI	MIFS
ALLAML	**1.25±0**	**1.25±3.95**	23.39±7.62	**1.25±4.52**	1.26±4.52	**1.25±3.95**
Carcinom	5.13±2.78	5.16±3.19	12.55±5.56	4.00±2.78	**5.10±2.62**	5.16±3.24
GLIOMA	14.00±0	36.00±9.66	32.00±11.35	**10.00±8.43**	16.00±6.32	**10.00±9.66**
lung	2.90±0	**1.98±3.15**	5.93±4.01	2.00±2.37	5.90±3.04	2.45±1.57
orlraws10P	1.00±4.22	7.00±6.75	8.00±6.66	3.00±4.22	1.00±3.16	**0.01±0**
pixraw10P	1.00±4.22	7.00±5.27	14.00±5.68	**0.01±3.16**	4.00±4.22	1.00±3.16
Prostate_GE	2.91±4.22	**1.82±4.18**	7.64±5.29	5.91±3.12	3.70±5.02	2.91±4.18
SMK_CAN_187	23.40±4.26	23.90±5.22	21.00±4.69	**18.70±4.70**	25.10±5.30	19.30±4.72
warpAR10P	13.80±6.07	21.50±6.28	36.15±6.74	9.20±3.24	5.40±4.37	**3.80±3.97**
warpPIE10P	3.33±2.30	**0.48±2.30**	3.40±0	1.90±2.30	0.95±2.30	2.38±2.46
Yale	51.50±7.22	47.30±4.26	48.05±6.20	38.27±6.20	**25.00±6.20**	40.70±6.99
Average Error	10.93	13.94	19.28	8.57	8.49	**8.09**
W/T/L	7/3/1	6/2/3	11/0/0	3/2/6	8/0/3	

Table 5 Classification Error Rate on Decision Tree Classifier

表 5 决策树分类器上的分类错误率 %

Dataset	CMIM	ICAP	CIFE	mRMR	L1MI	MIFS
ALLAML	8.40±6.63	6.80±7.10	22.20±7.10	1.40±5.88	1.40±4.45	**1.30±5.96**
Carcinom	6.40±4.09	6.75±5.75	18.40±6.86	**5.10±5.01**	6.30±5.59	**5.10±5.19**
GLIOMA	18.00±4.09	32.00±8.43	22.00±10.54	16.00±10.54	20.00±1.40	**12.00±11.35**
lung	4.50±4.09	5.40±4.36	7.40±3.82	**4.40±3.46**	6.80±5.26	**4.40±4.04**
orlraws10P	3.00±4.22	9.00±6.99	15.00±7.07	3.00±4.83	**1.00±6.32**	**1.00±5.68**
pixraw10P	1.00±4.22	4.00±4.22	4.00±4.21	**0.01±3.16**	2.00±4.83	1.00±3.16
Prostate_GE	**5.90±4.64**	6.00±4.73	8.80±4.09	**5.90±4.93**	7.60±6.42	6.90±4.73
SMK_CAN_187	26.30±4.75	27.90±5.21	29.40±4.56	**25.70±4.85**	32.00±5.54	27.20±4.36
warpAR10P	26.20±5.68	42.30±7.07	53.80±5.38	20.80±6.33	**11.50±6.49**	25.00±6.74
warpPIE10P	4.30±5.12	8.60±4.02	5.20±5.85	2.90±4.63	**1.90±5.24**	2.40±4.92
Yale	48.50±8.58	49.00±4.68	49.10±5.11	47.30±6.00	**36.60±6.31**	43.00±6.61
Average Error	13.86	17.98	21.39	12.05	**11.55**	11.75
W/T/L	8/1/2	10/0/1	11/0/0	5/2/4	7/1/3	

Table 6 Classification Error Rate on CNN Classifier

表 6 KNN 分类器上的分类错误率 %

Dataset	CMIM	ICAP	CIFE	mRMR	L1MI	MIFS
ALLAML	**2.68±5.96**	6.60±8.83	**2.68±7.18**	4.11±5.27	5.54±3.95	3.93±3.95
Carcinom	24.20±3.86	23.60±3.71	33.92±6.61	**15.00±3.24**	16.80±2.94	17.19±2.78
GLIOMA	20.00±3.86	20.00±11.35	22.00±10.54	**18.00±11.35**	24.00±10.54	**18.00±6.32**
lung	9.80±3.86	10.50±3.11	**5.93±4.74**	8.90±3.33	7.80±3.12	7.30±2.12
orlraws10P	**17.00±4.22**	28.00±6.67	33.00±4.71	20.00±4.22	13.00±3.16	20.00±0
pixraw10P	1.00±4.83	0.01±4.83	1.00±4.22	3.00±3.16	6.00±4.22	**0.00±0**
Prostate_GE	**5.90±4.14**	**5.90±5.40**	12.54±6.89	8.73±4.31	8.82±4.31	6.80±4.64
SMK_CAN_187	**23.50±4.68**	24.10±4.64	26.74±5.51	28.30±3.36	26.80±5.09	27.20±3.88
warpAR10P	40.80±6.07	48.50±6.33	58.46±6.74	24.60±5.38	**22.30±4.37**	26.20±7.65
warpPIE10P	19.05±2.01	27.62±2.01	26.03±3.17	15.24±2.30	21.43±2.46	**12.86±2.51**
Yale	61.40±6.30	51.80±6.43	51.73±5.46	48.00±7.94	45.30±7.65	**42.50±7.31**
Average Error	20.48	22.42	24.91	17.63	17.98	**16.54**
W/T/L	7/0/4	9/0/2	8/0/3	7/2/2	7/0/4	

从表 4 可以看出,在使用 SVM 作分类器时,MIFS 算法在 4 个数据集上取得了最低的分类错误率,并且平均分类错误率最低;在 W/T/L 指标中,MIFS 算法除了小幅落后于 mRMR 算法外,均优于其他方法. 因此 MIFS 算法选出的特征子集在 SVM 分类器上的表现良好.

从表 5 可以看出,在使用决策树作分类器时,MIFS 算法在 5 个数据集上取得了最低的分类错误率,接近全部数据集的一半. 对于平均分类错误率,MIFS 算法取得了第 2 名,仅与第 1 名相差 0.2 个百分点;在 W/T/L 指标中,MIFS 算法均优于其他方法.

从表 6 可以看出,MIFS 算法在 KNN 分类器上分别在 4 个数据集中取得了最低的分类错误率,同时取得了最低的平均分类错误率,低于第 2 名 1.09 个百分点;从 W/T/L 指标来看,MIFS 算法也都优于其他方法,因此 MIFS 算法选出的特征子集在 KNN 分类器上具有更好的分类性能.

综上,在使用最简单的支持向量机、决策树、KNN 三种分类器时,MIFS 方法都取得了很好的分类结果.

3 结　　论

本文提出了一种针对小样本数据的特征选择方法,该方法首先通过互信息对特征分组,选出组内与类别相关性最大的特征,大大降低了数据集的维度.同时为了解决无法自动给出最佳子集的问题,构造了过滤型与封装型算法结合的2阶段混合模型,即MIFS-Boruta算法,该算法不仅降低了数据集的维度,而且能够自动确定最佳特征子集,实验验证了所提算法的有效性.该算法为解决小样本问题提供了一种有效的方法.

然而,MIFS-Boruta算法的候选特征个数需要人为设定,如果设定的值过大,则会影响最终特征选择的运行效率;如果设定的值过小,则会影响最终选出特征的性能.因此,如何自动确定合理的候选特征个数还需要进一步的研究.

参　考　文　献

[1] Liu Bo, He Xiping. Feature Selection for High-Dimensional Data: Theory and Algorithm [M]. Beijing: Science Press, 2016 (in Chinese)
(刘波,何希平. 高维数据的特征选择:理论与算法[M]. 北京:科学出版社,2016)

[2] Liu Huan, Motoda H. Feature Selection for Knowledge Discovery and Data Mining [M]. New York: Kluwer Academic Publishers, 1998

[3] Nongnuch P. Practical approaches to mining of clinical datasets: From frameworks to novel feature selection [D]. Kingston Upon Hull, England: University of Hull, 2014

[4] Kira K, Rendell L A. The feature selection problem: Traditional methods and a new algorithm [C] //Proc of the 10th National Conf on Artificial Intelligence. Menlo Park: AAAI Press, 1992: 129-134

[5] Kononenko I. Estimation attributes: Analysis and extensions of RELIEF [C] //Proc of the 7th European Conf on Machine Learning. Berlin: Springer, 1994: 171-182

[6] Xu Lei, Yan Pingfan, Chang Tong. Best first strategy for feature selection [C] //Proc of the 9th Int Conf on Pattern Recognition. Piscataway, NJ: IEEE, 1988: 706-708

[7] Cai Zheyuan, Yu Jianguo, Li Xianpeng, et al. Feature selection algorithm based on kernel distance measure [J]. Pattern Recognition and Artificial Intelligence, 2010, 23(2): 235-240 (in Chinese)
(蔡哲元,余建国,李先鹏,等. 基于核空间距离测度的特征选择[J]. 模式识别与人工智能,2010,23(2):235-240)

[8] Jain A K, Robert P W, Mao Jianchang. Statistical pattern recognition: A review [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2000, 22(1): 4-37

[9] Xu Junling, Zhou Yuming, Chen Lin, et al. A unsupervised feature selection appprouch based on mutual information [J]. Journal of Computer Research and Development, 2012, 49(2): 372-382 (in Chinese)
(徐峻岭,周毓明,陈林,等. 基于互信息的无监督特征选择[J]. 计算机研究与发展,2012,49(2):372-382)

[10] Fleuret F. Fast binary feature selection with conditional mutual information [J]. Journal of Machine Learning Research, 2004, 5(3): 1531-1555

[11] Mitra P, Murthy C A, Sankar K P. Unsupervised feature selection using feature similarity [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2002, 24(3): 301-312

[12] Wei Hualiang, Billings S A. Feature subset selection and ranking for data dimensionality reduction [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2007, 29(1): 162-166

[13] Ding Chris, Peng Hanchuan. Minimum redundancy feature selection from microarray gene expression data [J]. Journal of Bioinformatics and Computational Biology, 2005, 3(2): 185-205

[14] Peng Hanchuan, Long Fuhui, Ding Chris. Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2005, 27(8): 1226-1238

[15] Hall M A. Correlation-based feature subset selection for machine learning [D]. Hamilton, NewZealand: University of Waikato, 1999

[16] Almuallim H, Dietterich T G. Learning with many irrelevant features [C] //Proc of the 9th National Conf on Artificial Intelligence. Menlo Park: AAAI Press, 1992: 547-552

[17] Liu Huan, Setiono R. A probabilistic approach to feature selection-A filter solution [C]//Proc of the 13th Int Conf on Machine Learning. New York: ILMS, 1996: 319-327

[18] Vergara J R, Estévez P A. A review of feature selection methods based on mutual information [J]. Neural Computing and Applications, 2014, 24(1): 175-186

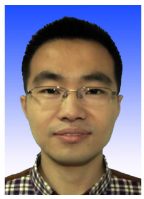
[19] Kursa M B, Rudnicki W R. Feature selection with the Boruta package [J]. Journal of Statistical Software, 2010, 36(11): 1-13

[20] Jakulin A, Bratko I. Testing the significance of attribute interactions [C] //Proc of the 21st Int Conf on Machine Learning. New York: ACM, 2004: 52-59

[21] Lin Dahua, Tang Xiaoou. An Integrated framework for feature extraction and fusion [C] //Proc of the 9th European Conf on Computer Vision. Berlin: Springer, 2006: 68-82

[22] Jitkrittum W, Hachiya H, Sugiyama M. Feature selection via L1-penalized squared-loss mutual information [J]. IEICE Trans on Information & Systems, 2013, E96-D(7): 1513-1524

Xu Hang, born in 1987. PhD candidate. Student member of CCF. Her main research interests include machine learning, data mining and service computing.



Zhang Kai, born in 1987. Master. His main research interests include machine learning and data mining (413804760@qq.com).

Wang Wenjian, born in 1968. PhD, professor, PhD supervisor. Senior member of CCF. Her main research interests include machine learning, data mining and computational intelligence.