DOI:10.11784/tdxbz201910030

基于对称不确定性和三路交互信息的特征子集选择算法

顾翔元，郭继昌，李重仪，肖利军
(天津大学电气自动化与信息工程学院，天津 300072)

摘　要：由于在评价冗余特征时只考虑对称不确定性或最大信息系数等某一种度量标准，使得现有的一些特征子集选择算法存在性能不理想的问题. 针对该问题，提出了一种基于对称不确定性和三路交互信息的特征子集选择算法. 首先，计算特征与类标签的对称不确定性，按照其值大小对特征作降序排序处理，并消除不相关特征；然后，计算特征间的对称不确定性以及特征与类标签的三路交互信息，并与特征与类标签的对称不确定性一起，经过比较和排序等运算以消除冗余特征而得到选取的特征. 在评价冗余特征上同时考虑对称不确定性和三路交互信息两种度量标准，并结合比较和排序等运算，可以减少将相关特征当作冗余特征而消除的情况，使得一些效果显著的相关特征得以保留. 为验证所提算法的性能，采用 J48、IB1 和 Naïve Bayes 3 种分类器将其与另外 4 种特征子集选择算法在 3 个 UCI 数据集和 9 个 ASU 数据集上进行实验. 实验结果表明，所提算法能够在选取特征数和用时均较少的情况下取得很好的特征选择效果.

关键词：特征子集选择；三路交互信息；对称不确定性；特征选择；排序
中图分类号：TP391　　**文献标志码**：A　　**文章编号**：0493-2137(2021)02-0214-07

Feature Subset Selection Algorithm Based on Symmetric Uncertainty and Three-Way Interaction Information

Gu Xiangyuan，Guo Jichang，Li Chongyi，Xiao Lijun
(School of Electrical and Information Engineering，Tianjin University，Tianjin 300072，China)

Abstract：It is known that only one metric is considered for evaluating redundant features such as symmetric uncertainty or maximum information coefficient and existing feature subset selection algorithms used for evaluation are not able to deliver the desired results. So our objective is to solve this problem and a feature subset selection algorithm based on symmetric uncertainty and three-way interaction information(SUTII) is proposed. First, symmetric uncertainty between features and the class label is evaluated, and features are arranged in descending order by ranking, and irrelevant features are removed. Then three-way interaction information among features and the class label and symmetric uncertainty between features are calculated and they are used jointly with symmetric uncertainty between features and the class label in a way of comparison and ranking calculation to remove redundant features. In this study, evaluating redundant features, both three-way interaction information and symmetric uncertainty are considered, and comparison and ranking calculation are adopted. The simulation that relevant feature are considered as redundant features and removed is decreased and some informative relevant features are retained. For validating the performance, SUTII is compared with four feature subset selection algorithms. Three classifiers J48, IB1, Naïve Bayes, three UCI datasets, and nine ASU datasets are used in the experiment. Experimental results demonstrate that SUTII can achieve better feature selection performance by means of few selected features and by consuming less time.

Keywords: feature subset selection; three-way interaction information; symmetric uncertainty; feature selection; ranking

收稿日期：2019-10-19；**修回日期**：2020-03-03.
作者简介：顾翔元(1990—　　)，男，博士研究生，gxiangyuan@tju.edu.cn.
通信作者：郭继昌，jcguo@tju.edu.cn.
基金项目：国家自然科学基金资助项目(61771334).
Supported by the National Natural Science Foundation of China(No. 61771334).

计算机技术的快速发展，使得各种数据量快速增长，智能信息处理所需数据集的特征维数也越来越高. 数据集特征维数的增加，会带来分类准确率下降、计算耗时多等问题. 因此有必要进行特征选择[1-4].

特征子集选择是特征选择的一种重要方法[5]，它利用某种度量标准对特征与类标签的相关性以及特征间的冗余性进行度量，消除不相关特征和冗余特征. 度量标准是影响特征子集选择算法效果好坏的关键.

文献[6-7]利用互信息实现了特征选择，互信息是一种常用的度量标准，它具有可描述特征间的非线性相关性和空间变换不变性等优点[8]，但其优先选取的特征不能保证取得较好的分类效果. 文献[9]对互信息进行了归一化处理，提出了对称不确定性(symmetric uncertainty，SU)度量标准. FCBF[10]和 FCBF-MIC 算法[11]利用对称不确定性进行了特征子集选择. FCBF 算法采用对称不确定性度量特征与类标签的相关性以及特征间的冗余性，并将特征与类标签的对称不确定性和特征间对称不确定性做比较来消除冗余特征. FCBF-MIC 算法分别采用对称不确定性和最大信息系数对特征与类标签的相关性以及特征间的冗余性进行度量，并利用特征与类标签的最大信息系数和特征间的最大信息系数的比较结果来消除冗余特征. 由于仅利用对称不确定性或最大信息系数等某一种度量标准来评价冗余特征，FCBF 和 FCBF-MIC 等算法存在将相关特征当作冗余特征而消除的问题.

针对上述问题，本文提出一种基于对称不确定性和三路交互信息的特征子集选择算法，首先利用对称不确定性进行相关性分析，消除不相关特征，然后分别利用对称不确定性和三路交互信息进行冗余性分析和交互性分析，消除冗余特征. 本算法由于利用了对称不确定性和三路交互信息两种度量标准而减少了相关特征的误消除，使得有效特征得以保留，因此算法获得了更好的性能.

1　三路交互信息和马尔可夫毯

1.1　三路交互信息

信息熵被用来量化某随机变量信息量的大小，其定义如式(1)所示.

$$H(X)=-\sum_{x\in X}p(x)\lg p(x) \tag{1}$$

式中 $p(x)$ 为一离散随机变量 X 取值为 x 的概率.

互信息被用来表述两变量所包含信息量的多少，两变量 X 和 Y 的互信息 $I(X;Y)$ 定义如式(2)所示.

$$I(X;Y)=\sum_{x\in X}\sum_{y\in Y}p(x,y)\lg\frac{p(x,y)}{p(x)p(y)} \tag{2}$$

式中：$p(x,y)$ 为两变量的联合概率密度函数；$p(y)$ 为变量 Y 取值为 y 的概率. $I(X;Y)$ 值越大，表明 X 和 Y 所包含的共同信息越多.

三路交互信息 $I(X;Y;Z)$ 是互信息的扩展，其可为正、负和零. 当 $I(X;Y;Z)$ 为正值时，表明两变量 X 和 Y 共同提供关于 Z 的信息要大于它们单独提供关于 Z 信息的和，此时表明 X 和 Y 在提供关于 Z 的信息上是互补的；其值为负值时，表明 X 和 Y 在提供关于 Z 的信息上是冗余的；其值为零值时，表明两变量在提供关于 Z 的信息上是独立的.

1.2　马尔可夫毯

马尔可夫毯可被用于特征选择中，其定义[10]如下.

定义 1(马尔可夫毯) 给定一个特征 f_j，类标签 c，一个特征集 F 和 F 中一个特征子集 G_j. 假设 $G_j\subset F(f_j\notin G_j)$，$G_j$ 为 f_j 的马尔可夫毯的条件是概率 $p(F-G_j-\{f_j\},c|f_j,G_j)=p(F-G-\{f_j\},c|G_j)$.

由定义 1 可知，如果选取 f_j，前后概率不变，则 G_j 为 f_j 的马尔可夫毯. 基于马尔可夫毯，对冗余特征有这样的定义：假设 F 是当前特征集，F 中的一个特征 f_j 是冗余特征当且仅当存在 $G_j\subset F(f_j\notin G_j)$ 为 f_j 的马尔可夫毯[10].

马尔可夫毯可以被用来评价冗余特征，然而由于运算量较大，其很少被使用，通常利用不同形式的近似马尔可夫毯. 如定义 2 所示，文献[10]给出一种近似马尔可夫毯来评价冗余特征.

定义 2(近似马尔可夫毯) 特征 f_s 是特征 f_i 的近似马尔可夫毯当且仅当同时满足 $\mathrm{SU}(c,f_s)\geqslant\mathrm{SU}(c,f_i)$ 和 $\mathrm{SU}(f_i,f_s)\geqslant\mathrm{SU}(c,f_i)$.

由定义 2 的 $\mathrm{SU}(f_i,f_s)\geqslant\mathrm{SU}(c,f_i)$ 可知，文献[10]将特征与类标签的对称不确定性和特征间的对称不确定性做比较，利用比较结果来评价冗余特征. 由于只考虑两个特征来评价冗余特征，使得满足 $\mathrm{SU}(f_i,f_s)\geqslant\mathrm{SU}(c,f_i)$ 条件的特征未必是冗余特征，会将一些相关特征当作冗余特征而消除.

2　本文算法

本文算法主要分为两个步骤：首先消除不相关特征，然后消除冗余特征，具体过程如图 1 所示.

图 1 中，首先计算特征与类标签的对称不确定性值，消除其值为零的特征；然后，计算特征间的对称

不确定性值,按其和特征与类标签的对称不确定性值的大小确定它们的序值,并计算特征与类标签的三路交互信息值. 若某特征满足以下条件:其与其他特征的对称不确定性值和序值分别大于其与类标签的对称不确定性值和序值,其与其他特征、类标签的三路交互信息值小于零,则将其作为冗余特征消除.

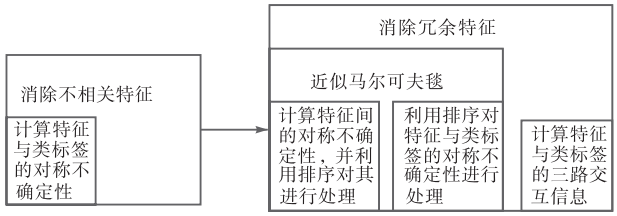

图 1 本文算法的框架

Fig.1 Framework of the proposed algorithm

2.1 相关特征的评价

相关性通常是指单个特征与类标签的关系. 利用对称不确定性进行相关性分析,特征 f_i 与类标签 c 的对称不确定性 $\mathrm{SU}(c,f_i)$ 如式(3)所示.

$$\mathrm{SU}(c,f_i)=\frac{2I(c;f_i)}{H(c)+H(f_i)} \tag{3}$$

式中 $\mathrm{SU}(c,f_i)\in[0,1]$. 与类标签的对称不确定性值越大,表明该特征与类标签越相关;当与类标签的对称不确定性值为零时,表明该特征与类标签是不相关的. 定义 3 给出相关特征的评价方法.

定义 3(相关特征的评价方法) f_i 是相关特征当且仅当满足 $\mathrm{SU}(c,f_i)>0$.

2.2 冗余特征的评价

利用定义 3 将不相关特征从原始特征集中消除后,需要消除冗余特征. 接下来进行冗余性分析.

冗余性通常是指特征间的关系. 利用对称不确定性进行冗余性分析.

鉴于文献[10]中近似马尔可夫毯存在的问题,分别利用排序对 $\mathrm{SU}(c,f_i)$ 和 $\mathrm{SU}(f_i,f_s)$ 进行处理,得到其序值 $R(c,f_i)$ 和 $R(f_i,f_s)$. 由于对称不确定性的最大值为 1,最小值为 0,而不同序值间最小的差别为 1. 所以,对于不同的 f_i,$R(c,f_i)$ 间的差别要大于 $\mathrm{SU}(c,f_i)$ 间的差别;对于同一个 f_i 和不同的 f_s,$R(f_i,f_s)$ 间的差别要大于 $\mathrm{SU}(f_i,f_s)$ 间的差别. 使得 $R(c,f_i)$ 和 $R(f_i,f_s)$ 的部分比较结果较 $\mathrm{SU}(c,f_i)$ 和 $\mathrm{SU}(f_i,f_s)$ 的比较结果有一定的优势.

如果只利用 $R(c,f_i)$ 和 $R(f_i,f_s)$ 的比较结果来定义近似马尔可夫毯,由于量级不对等,会存在一些比较结果不准确的问题.

而将 $\mathrm{SU}(c,f_i)$ 和 $\mathrm{SU}(f_i,f_s)$ 的比较结果与 $R(c,f_i)$ 和 $R(f_i,f_s)$ 的比较结果相结合来定义近似马尔可夫毯,不但可以减少将相关特征当作冗余特征而消除的情况,而且结果也更为准确. 所以,如定义 4 所示,给出一种近似马尔可夫毯.

定义 4(近似马尔可夫毯) f_s 是 f_i 的近似马尔可夫毯当且仅当同时满足 $\mathrm{SU}(c,f_s)>\mathrm{SU}(c,f_i)$、$\mathrm{SU}(f_i,f_s)>\mathrm{SU}(c,f_i)$ 和 $R(f_i,f_s)>R(c,f_i)$. 其中,$R(c,f_i)$ 和 $R(f_i,f_s)$ 分别是 $\mathrm{SU}(c,f_i)$ 和 $\mathrm{SU}(f_i,f_s)$ 经排序而得到的序值.

由定义 4 可知,同时满足 $\mathrm{SU}(f_i,f_s)>\mathrm{SU}(c,f_i)$ 和 $R(f_i,f_s)>R(c,f_i)$ 条件的特征才是冗余特征,可以保证效果较显著的特征被保留. 由于只考虑一种度量标准,会存在将相关特征当作冗余特征而消除的问题. 鉴于此,接着进行交互性分析.

交互性通常是指两个或多个特征与类标签的关系. 利用三路交互信息进行交互性分析.

由第 1.1 节可知,当 $I(f_i;f_s;c)<0$ 时,特征 f_i 与 f_s 共同提供关于类标签 c 的信息要小于它们单独提供关于 c 信息的和,可以表明 f_i 与 f_s 在提供关于 c 的信息上是冗余的. 因此,在定义 4 的基础上,引入三路交互信息,给出冗余特征的评价方法.

定义 5(冗余特征的评价方法) 候选特征 f_i 是冗余特征当且仅当同时满足 $\mathrm{SU}(f_i,f_s)>\mathrm{SU}(c,f_i)$、$R(f_i,f_s)>R(c,f_i)$ 和 $I(f_i;f_s;c)<0$.

在近似马尔可夫毯的基础上,引入三路交互信息来评价冗余特征,可以进一步减少将效果显著的相关特征当作冗余特征而消除的情况.

2.3 算法实现

利用定义 3 和定义 5,本文实现一种基于对称不确定性和三路交互信息的特征子集选择算法(简记为 SUTII),该算法的伪代码如下.

输入:M 数据集的特征数.
输出:S 已选特征集.
1:初始化 $S=\phi, Y=\phi, X=\{f_1,f_2,\cdots,f_M\}$;
2:for $f_i\in X$
3:计算 $\mathrm{SU}(c,f_i)$;
4:if $\mathrm{SU}(c,f_i)>\delta$
5:$Y=Y\cup\{f_i\}$;
6:end if
7:end for
8:按照 $\mathrm{SU}(c,f_i)$ 值,将 Y 中的特征做降序排序;
9:while $|Y|>1$
10:令 f_s 为 Y 的第 1 个特征;
11:$S=S\cup\{f_s\}$;
12:$Y=Y-\{f_s\}$;
13:for $f_i\in Y$

14：计算 $SU(c,f_i)$；
15：for $f_s \in S$
16：计算 $SU(f_i,f_s)$；
17：计算 $I(f_i;f_s;c)$；
18：end for
19：对 $SU(c,f_i)$排序，得到 $R(c,f_i)$；
20：对 $SU(f_i,f_s)$排序，得到 $R(f_i,f_s)$；
21：if $SU(f_i,f_s) > SU(c,f_i)$
22：if $R(f_i,f_s) > R(c,f_i)$
23：if $I(f_i;f_s;c) < 0$
24：$Y = Y - \{f_i\}$；
25：end if
26：end if
27：end if
28：end for
29：end while
30：if $|Y| = 1$
31：$S = S \bigcup Y$；
32：end if

SUTII 算法分为两部分：第 1 部分（第 1 ~ 8 行），先对候选特征集 X、已选特征集 S 和与 c 相关的特征集 Y 进行初始化，然后计算 X 中的特征与 c 的 SU，将与 c 相关的特征放入 Y 中，并将这些特征做降序排序；第 2 部分（第 9 ~ 32 行），先从 Y 中取出一个 SU 为最大值的特征，并放入 S 中；然后计算 f_i 与 S 中特征的 SU 和 c、f_i 与 S 中特征的三路交互信息，并分别对 $SU(c,f_i)$ 和 $SU(f_i,f_s)$ 进行排序处理，得到 $R(c,f_i)$ 和 $R(f_i,f_s)$；接着将 $SU(f_i,f_s)$ 与 $SU(c,f_i)$、$R(f_i,f_s)$ 与 $R(c,f_i)$ 以及 $I(f_i;f_s;c)$ 与零值做比较，如果 $SU(f_i,f_s)$ 大于 $SU(c,f_i)$、$R(f_i,f_s)$ 大于 $R(c,f_i)$ 和 $I(f_i;f_s;c)$ 小于零，将 f_i 从 Y 中消除. 按照上述步骤选取特征，直到 Y 中的特征少于 2 个结束. 最后，如果 Y 中有一个特征，将该特征放入 S 中.

3　实验结果

3.1　数据集和实验设置

表 1 给出用到的数据集，它们均取自 UCI 机器学习数据库[12]和 ASU 特征选择数据库[13]. 采用 J48、IB1 和 Naïve Bayes 分类器，其参数取 WEKA[14] 默认参数. 采用最小描述长度离散方法[15]. 特征选择过程用到 ASU 特征选择软件包[16]. 实验中，SUTII 算法的参数 δ 取 10^{-10}.

为减小随机性对最终结果的影响，进行 10 次十折交叉验证方法[17]处理，将 10 次结果的平均值作为最终结果，十折交叉验证是将数据集分成 10 等份，其中 9 份作为训练集，剩余的 1 份作为测试集，依次进行，直至所有的数据集都为测试集；此外，利用显著性水平为 5% 的单边配对样本 t 检验进行显著性检验.

表 1　实验中用到的数据集

Tab.1　Used datasets in the experiment

数据集	样本数	特征数	类	来源
Sonar	208	60	2	UCI
Musk	476	166	2	UCI
Mfeat_fac	2 000	216	10	UCI
Isolet	1 560	617	26	ASU
ORL	400	1 024	40	ASU
warpAR10P	130	2 400	10	ASU
lung	203	3 312	5	ASU
GLIOMA	50	4 434	4	ASU
arcene	200	10 000	2	ASU
pixraw10P	100	10 000	10	ASU
CLL_SUB_111	111	11 340	3	ASU
GLI_85	85	22 283	2	ASU

为验证 SUTII 算法的特征选择效果，将其与 FCBF-MIC[11]、SAOLA[6-7]、FCBF[10] 和 NFCBF 算法[18]做比较. 表 2 ~ 表 4 分别给出这些算法利用 J48、IB1 和 Naïve Bayes 分类器选取特征的分类准确率，W/T/L 行给出利用单边配对 t 检验而得到的值，其中，W、T 和 L 分别表示 SUTII 算法显著优于、无显著和显著劣于其他算法的数据集数. 表 5 给出这些算法选取的特征数，表 6 给出这些算法特征选择的用时.

3.2　实验结果与分析

由表 2 可知，SUTII 算法的平均值最大，而 FCBF-MIC 算法的平均值最小. 由 W/T/L 值可以得到，SUTII 算法优于 NFCBF、FCBF-MIC、SAOLA 和 FCBF 算法的数据集数分别为 4、10、2 和 2 个，而劣于这些算法的数据集数分别为 2、0、0 和 0 个，从而得知 SUTII 算法的特征选择效果优于另外 4 种算法.

表 3 表明，SUTII 算法的平均值最大. 由 W/T/L 值可知，SUTII 算法优于 NFCBF、FCBF-MIC、SAOLA 和 FCBF 算法的数据集数分别为 6、8、8 和 3 个，与表 2 相比，SUTII 算法较 SAOLA 和 FCBF 算法的优势有所增加.

表 4 中，NFCBF 算法的平均值较大. W/T/L 值表明 SUTII 算法的特征选择效果略优于 FCBF 算法、显著优于 FCBF-MIC 和 SAOLA 算法，而和 NFCBF 算法相当.

由表 5 的平均值可知，SAOLA 算法选取的特征最少，SUTII 算法选取的特征与 FCBF 算法相当，而多于 FCBF-MIC 和 SAOLA 算法，NFCBF 算法选取的特征最多，明显多于其他 4 种算法. SUTII 算法在

如 Sonar 和 Mfeat_fac 等数据集上选取的特征占数据集全部特征的比重较大，而在如 arcene、CLL_SUB_111 和 GLI_85 等数据集上选取的特征占的比重较小.

表 2 利用 J48 分类器选择特征的平均分类准确率

Tab.2 Average accuracy with J48 classifier on the selected features %

数据集	SUTII	NFCBF	FCBF-MIC	SAOLA	FCBF
Sonar	73.59	72.22	68.88	72.49	73.03
Musk	81.10	82.93	64.71	79.27	80.28
Mfeat_fac	88.75	88.80	88.01	88.43	88.77
Isolet	77.68	78.46	76.95	72.53	76.85
ORL	54.70	58.03	55.28	54.60	54.60
warpAR10P	69.15	70.00	62.69	66.69	68.69
lung	89.16	90.79	87.37	88.23	89.01
GLIOMA	67.20	59.60	48.60	68.20	64.40
arcene	76.25	71.95	68.30	76.60	77.10
pixraw10P	92.90	87.10	90.40	92.20	92.80
CLL_SUB_111	67.19	65.05	63.40	67.02	67.75
GLI_85	79.79	77.56	79.36	79.28	80.71
平均值	76.46	75.21	71.16	75.46	76.17
W/T/L		4/6/2	10/2/0	2/10/0	2/10/0

表 3 利用 IB1 分类器选择特征的平均分类准确率

Tab.3 Average accuracy with IB1 classifier on the selected features %

数据集	SUTII	NFCBF	FCBF-MIC	SAOLA	FCBF
Sonar	79.75	84.61	69.11	76.77	76.98
Musk	81.96	85.31	67.97	81.90	81.87
Mfeat_fac	96.31	96.27	95.97	96.04	96.30
Isolet	83.73	87.88	84.79	75.51	81.38
ORL	94.77	94.43	94.03	93.55	94.58
warpAR10P	80.15	40.77	66.08	76.00	77.08
lung	95.23	90.89	93.63	94.53	95.28
GLIOMA	75.20	66.20	69.40	72.00	72.80
arcene	80.85	83.10	68.10	76.25	79.85
pixraw10P	99.00	98.70	99.00	98.90	99.00
CLL_SUB_111	71.15	60.11	70.82	73.48	71.61
GLI_85	88.60	87.49	88.26	87.53	88.10
平均值	85.56	81.31	80.60	83.54	84.57
W/T/L		6/2/4	8/3/1	8/4/0	3/9/0

表 4 利用 Naïve Bayes 分类器选择特征的平均分类准确率

Tab.4 Average accuracy with Naïve Bayes classifier on the selected features %

数据集	SUTII	NFCBF	FCBF-MIC	SAOLA	FCBF
Sonar	73.48	72.28	72.30	72.87	73.36
Musk	77.43	82.14	64.89	75.62	76.50
Mfeat_fac	94.53	94.56	94.60	94.16	94.78
Isolet	84.94	90.31	86.43	77.50	82.50
ORL	88.30	88.75	83.78	84.90	86.92
warpAR10P	79.15	82.15	59.62	74.08	77.92
lung	96.15	95.11	94.38	95.91	95.96
GLIOMA	74.00	77.80	63.20	72.40	74.80
arcene	73.40	70.30	67.15	72.20	73.85
pixraw10P	100.00	99.40	97.30	99.80	99.90
CLL_SUB_111	79.52	77.95	71.74	78.67	79.36
GLI_85	83.76	84.13	85.96	83.67	82.58
平均值	83.72	84.57	78.45	81.81	83.20
W/T/L		4/4/4	8/2/2	5/7/0	2/9/1

表 5　各算法选择的特征数

Tab.5　Number of selected features of different algorithms

数据集	SUTII	NFCBF	FCBF-MIC	SAOLA	FCBF
Sonar	12.95	48.77	3.90	8.74	9.33
Musk	17.54	60.96	2.35	10.33	11.81
Mfeat_fac	54.31	59.46	32.77	34.60	39.60
Isolet	49.66	184.27	48.91	19.63	32.44
ORL	65.59	364.00	65.85	41.68	53.86
warpAR10P	37.56	1 847.63	10.40	15.65	22.44
lung	311.17	990.15	87.15	264.24	338.61
GLIOMA	50.39	2 047.72	95.11	23.66	43.98
arcene	36.33	7 413.04	5.19	17.68	22.18
pixraw10P	222.42	1 524.19	360.20	132.64	222.61
CLL_SUB_111	82.60	8 762.65	51.12	61.92	75.68
GLI_85	131.02	19 105.70	68.48	101.08	124.23
平均值	89.30	3 534.05	69.29	60.99	83.06

表 6 中，平均值表明 SUTII、SAOLA 和 FCBF 算法的特征选择用时较少，明显少于 NFCBF 和 FCBF-MIC 算法；NFCBF 算法在如 arcene、CLL_SUB_111 和 GLI_85 等数据集上的用时较多，而 FCBF-MIC 算法在如 Mfeat_fac、Isolet 和 ORL 等数据集上的用时较多.

表 6　各算法的用时

Tab.6　Time of different algorithms　　s

数据集	SUTII	NFCBF	FCBF-MIC	SAOLA	FCBF
Sonar	0.04	0.04	0.26	0.03	0.03
Musk	0.10	0.14	3.11	0.08	0.07
Mfeat_fac	0.97	0.46	289.76	0.48	0.43
Isolet	2.49	2.84	1 074.56	1.88	1.77
ORL	1.31	1.96	310.44	0.69	0.53
warpAR10P	0.54	21.03	11.30	0.41	0.37
lung	4.78	20.84	128.61	3.97	1.54
GLIOMA	0.82	43.31	8.14	0.76	0.54
arcene	1.63	405.92	17.49	1.53	1.37
pixraw10P	9.39	36.40	193.73	5.91	2.89
CLL_SUB_111	1.77	451.63	18.27	1.86	1.53
GLI_85	3.73	1 951.82	18.38	3.53	2.78
平均值	2.30	244.70	172.84	1.76	1.15

4　结　语

本文采用对称不确定性进行相关性分析和冗余性分析，利用三路交互信息进行交互性分析，提出一种基于对称不确定性和三路交互信息的特征子集选择算法 SUTII. 为验证 SUTII 算法的性能，利用 J48、IB1 和 Naïve Bayes 分类器将其与另外 4 种特征子集选择算法 NFCBF、FCBF-MIC、SAOLA 和 FCBF 在 3 个 UCI 数据集和 9 个 ASU 数据集上做对比实验. 实验结果表明，SUTII 算法能够取得较好的特征选择效果，同时选取的特征较 FCBF-MIC、SAOLA 和 FCBF 算法有所增加，验证了所提冗余特征的评价方法能够减少将相关特征当作冗余特征而消除的情况，使一些效果较显著的特征得以保留.

参考文献：

[1] Zhang R，Nie F P，Li X L，et al. Feature selection with multi-view data：A survey[J]. Information Fusion，2019，50：158-167.

[2] Liu C，Zheng C T，Wu S，et al. Multitask feature selection by graph-clustered feature sharing[J]. IEEE Transactions on Cybernetics，2020，50(1)：74-86.

[3] Wang H，Ling Z L，Y K，et al. Towards efficient and effective discovery of Markov blankets for feature selection[J]. Information Sciences，2020，509：227-242.

[4] 顾翔元，郭继昌，田煜衡，等. 基于条件互信息的空域隐写检测特征选择算法[J]. 天津大学学报：自然科学与工程技术版，2017，50(9)：961-966.
Gu Xiangyuan，Guo Jichang，Tian Yuheng，et al.

Spatial-domain steganalytic feature selection algorithm based on conditional mutual information[J]. Journal of Tianjin University: Science and Technology, 2017, 50(9): 961-966(in Chinese).

[5] Cai J, Luo J W, Wang S L, et al. Feature selection in machine learning: A new perspective[J]. Neurocomputing, 2018, 300: 70-79.

[6] Yu K, Wu X D, Ding W, et al. Towards scalable and accurate online feature selection for big data[C]// International Conference on Data Mining. Shenzhen, China, 2014: 660-669.

[7] Yu K, Wu X D, Ding W, et al. Scalable and accurate online feature selection for big data[J]. ACM Transactions on Knowledge Discovery from Data, 2016, 11(2): 1-39.

[8] Estevez P A, Tesmer M, Perez C A, et al. Normalized mutual information feature selection[J]. IEEE Transactions on Neural Networks, 2009, 20(2): 189-201.

[9] Press W H, Teukolsky S A, Vetterling W T, et al. Numerical Recipes in C[M]. Cambridge: Cambridge University Press, 1988.

[10] Yu L, Liu H. Efficient feature selection via analysis of relevance and redundancy[J]. Journal of Machine Learning Research, 2004, 5: 1205-1224.

[11] 孙广路, 宋智超, 刘金来, 等. 基于最大信息系数和近似马尔科夫毯的特征选择方法[J]. 自动化学报, 2017, 43(5): 795-805.
Sun Guanglu, Song Zhichao, Liu Jinlai, et al. Feature selection method based on maximum information coefficient and approximate markov blanket[J]. Acta Automatica Sinica, 2017, 43(5): 795-805(in Chinese).

[12] Dua D, Graff C. UCI Machine Learning Repository [EB/OL]. http://archive.ics.uci.edu/ml, 2019-01-06.

[13] Li J D, Cheng K W, Wang S H, et al. Feature selection: A data perspective[J]. ACM Computing Surveys, 2018, 50(6): 1-73.

[14] Hall M, Frank E, Holmes G, et al. The WEKA data mining software: An update[J]. SIGKDD Explorations, 2009, 11(1): 10-18.

[15] Fayyad U M, Irani K B. Multi-interval discretization of continuous-valued attributes for classification learning [C]// International Joint Conference on Artificial Intelligence. Chambèry, France, 1993: 1022-1027.

[16] Zhao Z, Morstatter F, Sharma S, et al. ASU Feature Selection Software Package[EB/OL]. http://featureselection.asu.edu, 2010-12-16.

[17] Gu X Y, Guo J C, Xiao L J, et al. A feature selection algorithm based on equal interval division and minimal-redundancy-maximal-relevance[J]. Neural Processing Letters, 2020, 51(2): 1237-1263.

[18] 张 俐, 袁玉宇, 王 枞, 等. 基于最大相关信息系数的 FCBF 特征选择算法[J]. 北京邮电大学学报, 2018, 41(4): 86-90.
Zhang Li, Yuan Yuyu, Wang Cong, et al. FCBF feature selection algorithm based on maximum information coefficient[J]. Journal of Beijing University of Posts and Telecommunications, 2018, 41(4): 86-90(in Chinese).

(责任编辑: 王晓燕)