文章编号：2095-2163(2023)05-0131-09 中图分类号：TP181 文献标志码：A

医疗诊断上一种基于特征交互的 MIFS 算法

王新利[1]，李雨沛[1]，李海洋[2]
（1 上海理工大学 理学院，上海 200093；2 西交利物浦大学 智能工程学院，江苏 苏州 215000）

摘 要：MIFS 算法及其改进算法对医疗诊断数据集进行特征选择时，秉承“最大相关最小冗余”的思想，关注特征与类别的相关信息和特征之间的冗余信息，没有考虑到特征之间的交互信息。考虑到医疗诊断指标之间的交互，本文提出一种基于特征交互的 MIFS 算法（Feature Interaction Based MIFS Algorithm，MIFS-FI），在实现“最大相关”的同时，最大程度地去除冗余特征，保留交互特征，还有效地解决了 MIFS 算法中参数不确定以及相关项与冗余项不可比的问题。将 MIFS-FI 算法和其他 7 种基于互信息的特征选择方法应用于 14 个医疗诊断数据集进行对比实验，结果表明 MIFS-FI 算法在分类准确率、召回率和 $F1$ 值三方面优于其他 7 种特征选择方法，提高了分类精度。

关键词：特征选择；交互信息；冗余特征；分类精度

A feature interaction based MIFS algorithm for medical diagnosis

WANG Xinli[1], LI Yupei[1], LI Haiyang[2]
(1 College of Science, University of Shanghai for Science and Technology, Shanghai 200093, China;
2 School of Advanced Technology, Xi'an Jiaotong-Liverpool University, Suzhou Jiangsu 215000, China)

【Abstract】 MIFS algorithm and its improved algorithms adhere to the idea of "maximum correlation and minimum redundancy" to select features in medical diagnostic data sets, which pay attention to the relevant information and the redundant information, but do not consider the interaction information. In order to emphasize the role of Interaction information, Feature Interaction Based MIFS Algorithm (MIFS-FI) is proposed. MIFS-FI algorithm achieve "maximum correlation", and the redundant features are almost removed and the interactive features are nearly retained. Secondly, it effectively solves the problems of parameter uncertainty and correlation- redundancy incomparable in MIFS algorithm. Finally, the MIFS-FI algorithm and seven other feature selection methods based on mutual information are compared to 14 medical diagnosis datasets, and the results show that the MIFS-FI algorithm outperforms the others in terms of classification accuracy, recall, F_1 score and classification accuracy.

【Key words】 feature selection; interactive information; redundancy feature; classification accuracy

0 引 言

近年来，随着医疗诊断数据增多，检测指标也随之增加，如何从大批量的诊断指标中筛选出对诊疗判断最为有利的指标，是机器学习在医疗领域应用中的一个研究热点[1]。

现有的研究通常将诊断的指标看作特征，患病的程度看作类别，筛选诊断指标，即选择对判断类别有利的特征，去除与判断类别无关的指标，也就是选择“好的”特征，去除“坏的”特征。特征选择主要有 3 种常用的方法，分别是包裹法、嵌入法和过滤法。与过滤法相比，包裹法和嵌入法的特征选择效果更好，但是存在过拟合、计算复杂度高和效率低等问题，而过滤法的评价准则简单、运算效率高，应用范围更广泛[2]。

过滤法根据不同的评价准则来选择最优的特征子集，常用的评价准则有距离度量标准、一致性度量标准、依赖性度量标准和信息度量标准等[3]。距离度量标准用几何距离或者概率距离的大小来度量特征；一致性度量标准根据不一致样本数与总体样本比率来评估特征；依赖性度量标准则根据特征与类别的相关系数和特征之间的冗余性来判断特征；信息度量标准通过熵、互信息以及交互信息等来评价特征。相比于其他度量标准，信息度量标准可以衡量特征之间、特征与类别之间的非线性关系，因此信息度量标准作为过滤法的特征选择准则被广泛应

基金项目：国家自然科学基金(62073223)。

作者简介：王新利(1975-)，女，博士，讲师，主要研究方向：数据挖掘；李雨沛(1997-)，女，硕士研究生，主要研究方向：数据挖掘。
通讯作者：王新利 Email：xlwang602@163.com

收稿日期：2022-12-27

用[4]。在信息度量标准中,“好的”特征指与某类别互信息大的特征,“坏的”特征指与已选特征的互信息大的特征。互信息可以度量特征与类别的相关性或两个特征之间的相关性,互信息越大,则相关性越大。特别地,两个特征之间的“相关”称为“冗余”,两者之间的互信息越大则冗余性越大。

另一方面,交互信息也是信息度量标准中的一个重要的评价指标。例如,在异或问题中,两个特征分别与类别无关,但是这两个特征联合起来与类别有强相关性,说明这两个特征有交互性,称这两个特征为交互特征。在许多特征选择的算法设计中,交互信息也越来越被重视。姜文煊等[5]将交互信息加入到基于互信息的评价指标中,得到一种新的评价标准,并将其应用于地质评价中,获得了较高的评价准确率;陈昊楠等[6]根据交互信息选择交互特征,根据条件互信息最大化选择低冗余的特征,将两者结合得到一种新的特征选择方法,并应用于癌症分类中,有效提高了分类的准确率;顾翔元等[7]使用对称不确定性计算特征的相关性,再计算特征的交互信息来消除冗余特征,在不同的分类器上都获得了较高的精度。

根据信息度量标准的过滤法进行特征选择时,互信息特征选择算法(MIFS)是一种经典的算法,其根据特征与类别之间的互信息来衡量二者的相关,用特征之间的互信息来衡量两个特征的冗余,通过参数来调整去除冗余的大小[8]。MIFS 算法可以有效地选出与类别相关性大,特征之间冗余性小的特征,但是随着已选特征数量增加,冗余信息也会随之增多,进而增大与相关信息的差值,导致算法过度重视“冗余”而忽略“相关”。为解决这个问题,Kwak 等[9]在 MIFS 算法的基础上,增加了一个系数,用来平衡相关信息与冗余信息不可比的情况,得到了 MIFS-U 算法。MIFS 算法和 MIFS-U 算法都含有参数,参数的不确定性导致了这两个算法自适应性不强。进一步地,Peng 等[10]提出了一种不含参数的最小冗余最大相关算法(mRMR),用已选特征子集个数的倒数来代替参数,使算法具有更强的普适性和自适应性;ESTÉVEZ 等[11]在 mRMR 算法的基础上,将冗余信息的取值范围控制在 0 到 1 之间,对冗余项做归一化处理,得到了标准化互信息特征选择方法(NMIFS);Zhang 等[12]提出一种权重系数的加权归一化信息过滤准则,进一步解决了相关信息和冗余信息不平衡的问题。

MIFS 算法以及其各种改进算法都是围绕着“度量冗余”做改进,应用于医疗临床诊断数据分类时,分类效果较好,但是其并未考虑不同诊断指标之间的交互性。为进一步提高医疗临床诊断的分类精度,本文在 MIFS 算法的基础上提出一种基于特征交互的 MIFS 算法,应用于医疗临床诊断数据的分类。首先,根据特征交互信息和冗余信息的关系,重新定义不含参数的冗余系数,最大程度保留特征的交互信息,去除冗余信息;其次,用已选特征子集个数的倒数来平衡相关信息与冗余信息的不可比;最后,与其他 7 种基于互信息的特征选择方法比较,证明该算法精度明显高于其他方法。

1 背景知识

信息论中,通常用信息熵和互信息来度量特征和类别的相关性、特征之间的冗余性[13]。特征 f_i 的信息熵(Entropy)定义如式(1):

$$H(f_i) = -\sum p(f_i) \cdot \log(p(f_i)) \tag{1}$$

其中,$p(f_i)$ 表示特征 f_i 的概率密度函数,$H(f_i)$ 的取值在 0~1 之间。

特征 f_i 和 f_j 的联合熵(Joint Entropy)和条件熵(Conditional Entropy)定义如式(2)和式(3):

$$H(f_i,f_j) = -\sum\sum p(f_i,f_j) \cdot \log(p(f_i,f_j)) \tag{2}$$

$$H(f_i \mid f_j) = -\sum\sum p(f_i,f_j) \cdot \log(p(f_i \mid f_j)) \tag{3}$$

其中,$p(f_i,f_j)$ 表示特征 f_i 和 f_j 的联合概率密度函数,$p(f_i \mid f_j)$ 表示在特征 f_j 的条件下 f_i 的概率密度函数。

特征 $f_1,\cdots,f_n$ 的联合熵定义如式(4):

$$H(f_1,f_2,\cdots,f_n) = -\sum\cdots\sum p(f_1,f_2,\cdots,f_n) \cdot \log(p(f_1,f_2,\cdots,f_n)) \tag{4}$$

其中,$p(f_1,f_2,\cdots,f_n)$ 表示特征 $f_1,\cdots,f_n$ 的联合概率密度函数。

特征 f_i 和 f_j 的互信息(Mutual Information, MI)定义如式(5):

$$I(f_i;f_j) = \sum\sum p(f_i,f_j) \cdot \log\frac{p(f_i,f_j)}{p(f_i) \cdot p(f_j)} \tag{5}$$

互信息可以度量特征与类别或特征之间的相关性,当一个特征与类别的互信息越大时,这个特征与类别之间的相关度越大;当两个特征之间的互信息越大时,则这两个特征的冗余度越大。

熵、互信息之间的关系如式(6)和式(7):

$$H(f_i,f_j) = H(f_i) + H(f_j) - I(f_i;f_j) \tag{6}$$

$$I(f_i;f_j) = H(f_i) - H(f_i \mid f_j) \tag{7}$$

除了熵和互信息，交互信息也是信息论中重要的度量指标，交互信息又称为交互增益（Interaction Gain，IG），指的是三方或者多方的交互作用，通常三方的交互是指特征 f_i 和 f_j 以及类别 C 之间的交互信息，多方则是多个特征之间与类别的交互信息。三方交互增益的定义如式（8）[14]：

$$IG(f_i;f_j;C) = I(f_i;C) + I(f_j;C) + I(f_i;f_j) + H(f_i,f_j,C) - H(C) - H(f_i) - H(f_j) \quad (8)$$

其中，$I(f_i;C)$ 表示特征 f_i 和类别 C 的互信息；$I(f_j;C)$ 表示特征 f_j 和类别 C 的互信息；$H(f_i,f_j,C)$ 是特征 f_i 和 f_j 以及类别 C 的联合熵；$H(C)$ 是类别 C 的熵；$H(f_i)$ 是特征 f_i 的熵；$H(f_j)$ 是特征 f_j 的熵。

根据熵与互信息的定义，交互信息的定义还可以用式（9）表示：

$$IG(f_i;f_j;C) = I(f_i,f_j;C) - I(f_i;C) - I(f_j;C) \quad (9)$$

其中，$I(f_i,f_j;C)$ 表示特征 f_i 和 f_j 与类别 C 的联合互信息。

当 $IG(f_i;f_j;C) < 0$ 或者 $IG(f_i;f_j;C) = 0$ 时，说明特征 f_i 和 f_j 与类别无关或者两者提供了相似信息；当 $IG(f_i;f_j;C) > 0$ 时，表示特征 f_i 和 f_j 组合提供的信息量大于特征 f_i 和 f_j 分别提供的信息量之和，说明特征 f_i 与 f_j 具有交互性。

2　相关工作

下面介绍一些经典的基于互信息的特征选择算法，其中 C 表示类别，F 表示原始特征集，S 表示已选特征集，$f_j \in F$ 表示候选特征，$f_i \in S$ 表示已选特征。

Battit 等[8]提出了基于互信息的特征选择算法（Mutual Information Feature Selection，MIFS）。该算法通过最大化特征与类别之间的互信息，最小化特征之间的互信息来选择特征。MIFS 算法的评价准则如式（10）：

$$\text{MIFS} = \underset{f_j}{\text{argmax}}(I(f_j;C) - \alpha \sum_{f_i \in S} I(f_i;f_j)) \quad (10)$$

其中，$I(f_j;C)$ 表示候选特征 f_j 与类别 C 的相关信息；$I(f_i;f_j)$ 表示已选特征 f_i 与候选特征 f_j 的冗余信息；参数 α 表示冗余系数，范围在 0~1 之间，当参数为 0 时，算法只计算相关信息，完全忽略冗余信息。

MIFS 算法可以有效地选出与类别相关性大，特征之间冗余性小的特征。但是当已选特征的数量变多时，冗余项相对于相关项会变得很大，这两项可能不在一个数量级上，导致冗余项占主导地位，相关项可能被忽略。

为了解决相关项与冗余项不平衡的问题，Kwak 等[9]在冗余项中加入系数 $\frac{I(f_j;C)}{H(f_j)}$ 来平衡两项不可比，提出了一致性分布的互信息特征选择方法（Mutual Information Feature Selection Under Uniform Information Distribution，MIFS-U），评价准则如式（11）：

$$\text{MIFS-U} = \underset{f_j}{\text{argmax}}(I(f_j;C) - \alpha \sum_{f_i \in S} \frac{I(f_j;C)}{H(f_j)} \times I(f_i;f_j)) \quad (11)$$

MIFS-U 算法在一定程度上缓解了相关项与冗余项的不平衡问题，但是在 MIFS 和 MIFS-U 算法的评价准则中都有需要调节的参数，取值具有一定的随机性和主观性，导致算法的自适应性不强。在 MIFS 算法的基础上，Peng 等[10]提出了一种不含参数的最大相关最小冗余算法（Minimal Redundancy Maximum Relevance，mRMR），评价准则如式（12）所示：

$$\text{mRMR} = \underset{f_j}{\text{argmax}}(I(f_j;C) - \frac{1}{|S|} \sum_{f_i \in S} I(f_i;f_j)) \quad (12)$$

该方法用已选子集个数的倒数来代替参数 α，不仅解决相关项与冗余项不平衡的问题，还使算法更具自适应性。在算法 MIFS、MIFS-U 和 mRMR 的评价准则中，都只涉及了特征的相关信息和冗余信息，特征和类别之间的交互性并未考虑。

Bennasar 等[15]关注到了交互信息的重要性，提出一种特征交互最大化的特征选择方法（Feature Interaction Maximization，FIM），其评价准则为式（13）：

$$\text{FIM} = \underset{f_j}{\text{argmax}}(I(f_j;C) + \min_{f_i \in S}(IG(f_i;f_j;C))) \quad (13)$$

Salem 等[16]注意到粗糙邻域集中的特征交互信息，并根据模糊联合互信息最大化的原则来选择特征，提出了基于模糊联合互信息最大化的特征选择方法（Feature Selection based on Fuzzy Joint Mutual Information Maximization，FJMJM），其评价准则如式（14）：

$$\text{FJMIM} = \underset{f_j}{\text{argmax}}(\min_{f_i \in S}(I(f_j,f_i;C) - IG(f_j;f_i;C))) \quad (14)$$

Wan 等[17]提出一种混合式特征选择方法来尽可能地保留交互特征，去除冗余特征；Gu 等[18]重视三方交互信息在特征选择中的作用，提出了一种基于等间隔划分和三方交互信息的特征子集选择算法来优化特征选择的效果，提高算法精度。虽然这些

算法也获得了较好的特征选择结果,但在重视交互的情况下,对冗余的关注却被大大降低。如何同时关注特征相关、冗余和交互,最大程度地保留相关、交互的特征,去除冗余特征,这是基于互信息的特征选择算法改进的一个重要的研究内容。

3 基于特征交互的 MIFS 算法

MIFS 算法及其改进算法采用最大相关最小冗余评价准则,可以较有效地选出特征子集。两特征与类别之间的互信息、交互信息、熵的关系如图 1 所示。图 1 中 a 部分表示在类别 C 下,特征 f_i 和 f_j 的互信息 $I(f_i;f_j \mid C)$,可以用公式(15)表示;b 部分表示两特征与类别之间的交互信息 $IG(f_i;f_j;C)$;c 部分表示在特征 f_j 的条件下,特征 f_i 和类别 C 的互信息 $I(f_i;C \mid f_j)$;d 部分表示在特征 f_i 条件下,特征 f_j 和类别 C 的互信息 $I(f_j;C \mid f_i)$。

$$I(f_i;f_j \mid C) = H(f_i \mid C) - H(f_i \mid f_j,C) \quad (15)$$

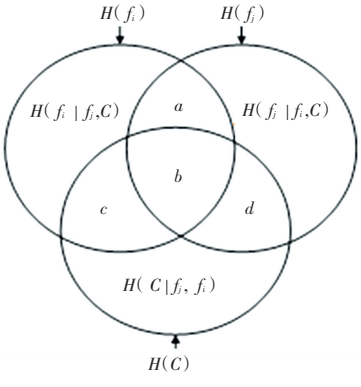


图 1 两特征与类别之间的互信息、交互信息、熵的关系

Fig. 1 Relation of mutual information, interactive information and entropy between the two features and categories

MIFS 算法在实现“最小冗余”的目标时,同时产生了“最小交互”。然而,当交互信息越大时,特征越应该被选入特征子集,因此需要最大程度地保留交互信息,以“最大交互”为目标。为了实现特征和类别的最大相关,同时尽可能兼顾特征之间的最小冗余和最大交互,本文提出了一种基于特征交互的 MIFS 算法(Feature Interaction Based MIFS Algorithm, MIFS-FI),评价准则如式(16):

$$\text{MIFS-FI} = \underset{f_j \in F-S}{\operatorname{argmax}}\left(I(f_j;C) - \frac{1}{|S|}\sum_{f_i \in S}\frac{I(f_j;f_i \mid C)}{IG(f_j;f_i;C)} \times I(f_j;f_i)\right), IG(f_j;f_i;C) \neq 0 \quad (16)$$

其中,S 表示已选特征子集;F 表示原始特征集;f_i 表示已选特征;f_j 表示候选特征;$I(f_j;f_i)$ 表示特征 f_i 和 f_j 的互信息;$I(f_j;C)$ 表示特征 f_j 和类别 C 的互信息;$I(f_j;f_i \mid C)$ 表示在类别 C 的条件下,特征 f_i 和 f_j 的互信息;$IG(f_j;f_i;C)$ 表示特征 f_j 和 f_i 以及类别 C 的三方交互信息。

在式(16)中,第一项是相关项,表示特征 f_j 和类别 C 的相关信息;第二项是冗余项,在冗余项中有两个系数,分别是 $\frac{I(f_j;f_i \mid C)}{IG(f_j;f_i;C)}$ 和 $\frac{1}{|S|}$。对于系数 $\frac{I(f_j;f_i \mid C)}{IG(f_j;f_i;C)}$,当 $IG(f_j;f_i;C) < 0$ 时,说明特征 f_j 与类别无关或者提供了与特征 f_i 相似的信息,在算法的设计中选择将特征 f_j 去除,所以仅需考虑 $IG(f_j;f_i;C) > 0$。考虑到,$I(f_j;f_i \mid C)$ 和 $IG(f_j;f_i;C)$ 合起来构成冗余信息,因此,当 $\frac{I(f_j;f_i \mid C)}{IG(f_j;f_i;C)} < 1$ 时,交互信息在冗余信息中占比越大,则去除的冗余信息越少,更偏向于“最大交互”,如图 2 所示;当 $\frac{I(f_j;f_i \mid C)}{IG(f_j;f_i;C)} > 1$ 时,交互信息在冗余信息中占比越小,则去除的冗余信息越多,更偏向于“最小冗余”,情况,如图 3 所示;当 $\frac{I(f_j;f_i \mid C)}{IG(f_j;f_i;C)} = 1$ 时,MIFS-FI 算法即 mRMR 算法。另一方面,冗余项是已选特征与候选特征互信息的累加和,当已选特征的数量增多时,冗余项会远远大于相关项,导致冗余项与相关项不可比,通过在冗余项中添加系数 $\frac{1}{|S|}$,可以在一定程度上缓解这个问题。

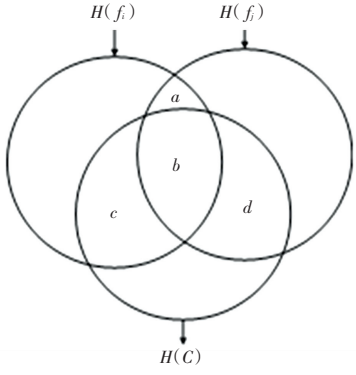


图 2 $\frac{I(f_j;f_i \mid C)}{IG(f_j;f_i;C)}<1$ 时特征与类别的关系

Fig. 2 Relationship between features and categories when $\frac{I(f_j;f_i \mid C)}{IG(f_j;f_i;C)}<1$

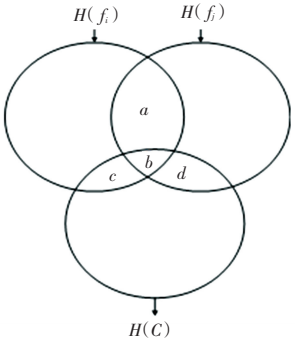


图 3　$\frac{I(f_j;f_i|C)}{IG(f_j;f_i;C)}>1$ 特征与类别的关系

Fig. 3　Relationship between features and categories when $\frac{I(f_j;f_i|C)}{IG(f_j;f_i;C)}>1$

基于特征交互的 MIFS 算法(MIFS-FI)具体流程见表 1。

表 1　MIFS-FI 算法流程

Tab. 1　MIFS-FI algorithm flow

MIFS-FI 算法流程
输入:特征集 $F=\{f_1,f_2,\cdots,f_n\}$,类别集 $C=\{C_1,C_2,\cdots,C_m\}$,阈值 k
输出:特征子集 S
1.初始化特征子集 $S=\varnothing$;
2.For $i=0$ to n
3. 计算所有特征与类别之间的互信息 $I(f_i;C)$,$i=1\cdots n$
4.End For
5.选出与类别互信息最大的特征 f_{k_1} 放入特征子集 S 中,将 F 中的特征 f_{k_1} 剔除
6.While $\|S\|<k$
7.　For $i=0$ to $\|F\|$
8.计算 F 中每一个候选特征 f_j 与已选特征 f_i 的交互信息 $IG(f_j;f_i;C)$
9.If $IG(f_j;f_i;C)<0$
10. $F=F-\{f_j\}$
11.　Else
12.　用公式(17)计算每一个候选特征 $f_j\in F$ 的特征得分
13.　End If
14.　End For
15.　选出特征得分最高的特征 f_{k_2} 加入特征子集 S 中,并在特征集 F 中删除此特征
16.End While

4　实　验

4.1　数据集与数据集的处理

本文选取 14 个关于医疗诊断的数据集来验证本文所提出算法的有效性。除了第 7 个数据集均来自 Matlab 数据库,其他 13 个实验数据集来自美国加州大学欧文分校提供的 UCI 数据库,14 个数据集的样本个数、特征数和类别个数见表 2。

表 2　数据集的描述

Tab. 2　Description of the dataset

数据集	样本数	特征数	类别数	记作
WDBC	569	32	2	D_1
Lung Cancer	32	56	2	D_2
Heart Disease	270	13	2	D_3
Cervical cancer	858	36	2	D_4
Lymphography	148	19	4	D_5
HCV	615	11	4	D_6
High-resolution ovarian cancer	216	4 000	2	D_7
Arrhythmia	452	279	16	D_8
Bone marrow transplant: children	187	39	2	D_9
Dermatology	366	33	6	D_{10}
Hepatitis	155	19	2	D_{11}
Horse Colic	368	27	2	D_{12}
Primary Tumor	339	17	3	D_{13}
Risk Factor prediction of Chronic Kidney Disease	202	28	2	D_{14}

表 2 中的数据集,有些存在不同程度的特征值缺失,本文采用均值替代法对存在缺失值的数据集进行填补后做归一化处理,x_{inew} 表示 x_i 归一化之后的样本,式(17):

$$x_{inew}=\frac{x_i-x_{\min}}{x_{\max}-x_{\min}} \tag{17}$$

其中,x_i 表示第 i 个样本;$x_{\min}$ 表示样本中的最小值;$x_{\max}$ 表示样本中的最大值。

归一化数据有利于加快模型的收敛速度。

4.2　对比方法介绍和算法评价指标

为验证本文提出算法(MIFS-FI)的有效性,与 7 种基于互信息的特征选择方法进行对比实验。这 7 种方法分别是互信息最大特征选择算法(Mutual Information Maximum, MIM)、基于互信息的特征选择算法(Mutual Information Feature Selection, MIFS)、最大相关最小冗余算法(Minimal Redundancy Maximum Relevance, mRMR)、条件信息特征提取算法(Conditional Informative Feature Extraction, CIFE)、基于模糊联合互信息最大化的特征选择方法(Feature Selection based on Fuzzy Joint Mutual Information Maximization, FJMJM)、动态变化特征选择算法(Dynamic Change of Selected Feature, DCSF)和特征交互最大化的特征选择方法(Feature Interaction Maximization, FIM)。

其中 MIM 算法只考虑了特征相关的算法,MIFS 算法、mRMR 算法、CIFE 算法、DCSF 算法考虑了特

征的相关和冗余。另外,MIFS-FI 算法重视了交互信息,需要选取一些关注了交互信息的算法进行对比实验。本文选取考虑了特征交互和相关的 FJMJM 算法、FIM 算法进行对比,上述几种算法的构造见表 3。

表 3 算法构造比较

Tab. 3 Comparison of algorithm construction

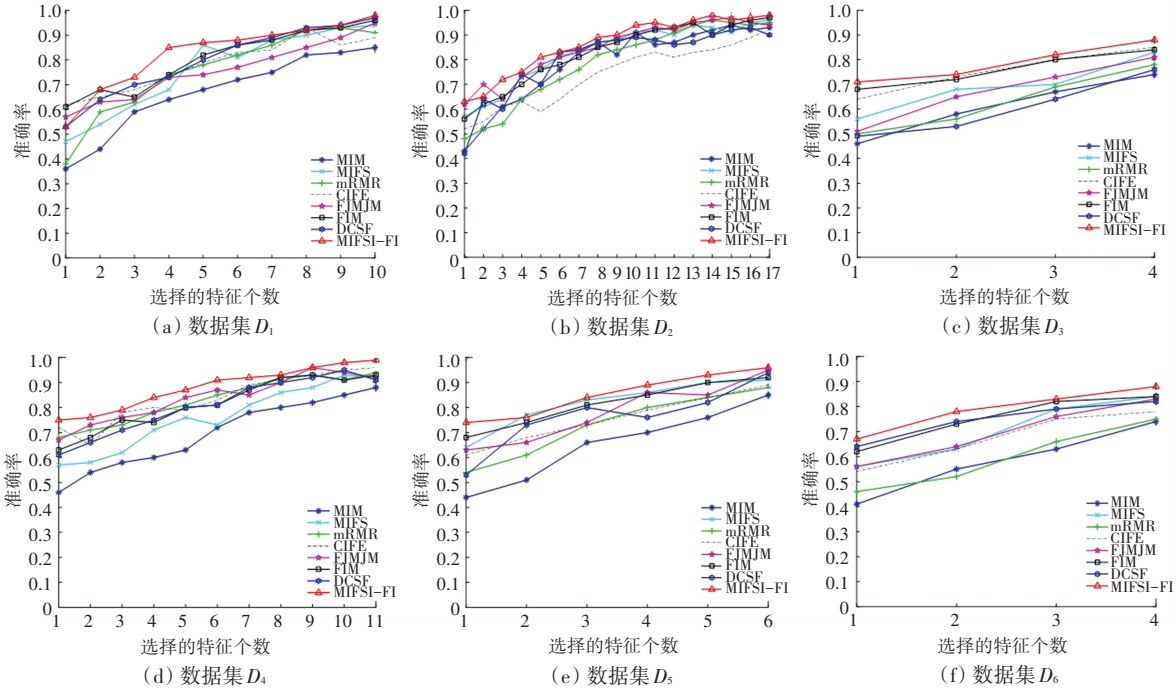
算法	是否不含参数	是否同时有相关项和冗余项	是否平衡相关项和冗余项	是否重视交互信息
MIM 算法	是	否	—	否
MIFS 算法	否	是	否	否
mRMR 算法	是	是	是	否
CIFE 算法	是	是	否	否
DCSF 算法	是	是	是	否
FIM 算法	是	否	—	是
FJMJM 算法	是	否	—	是
MIFS-FI 算法	是	是	是	是

由于 BP 神经网络分类精度高,且具有强自适应性、非线性映射等优点,被广泛应用于医疗诊断分类,因此本文选用 BP 神经网络模型做分类器来检验选择特征的质量,其评价标准为分类准确率(ACC)、F_1 指数和召回率。

4.3 实验结果与分析

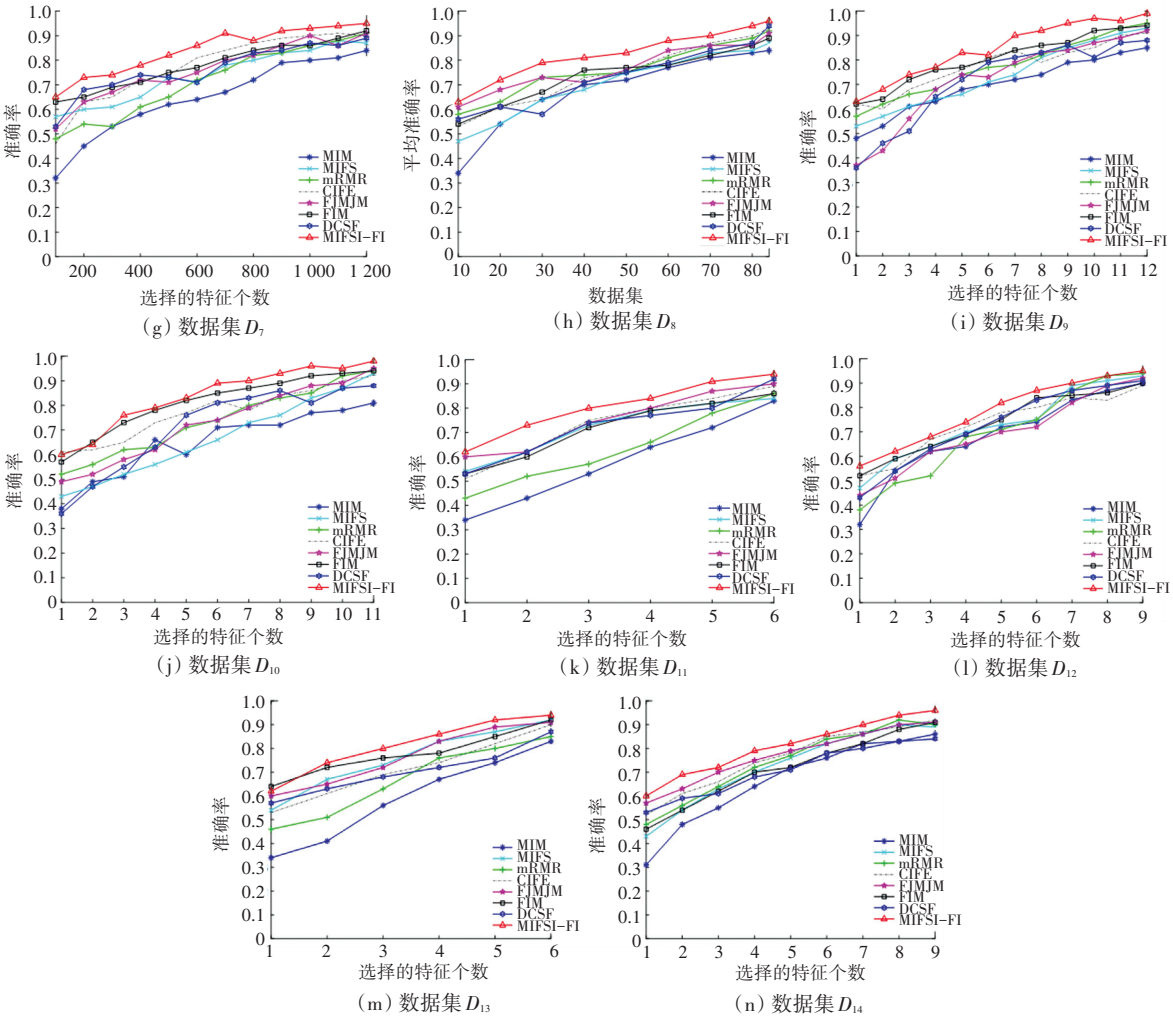
实验中所有特征选择方法选择特征数量不超过总特征的 30%,BP 神经网络的迭代次数设置为 1 000,学习率设置为 0.02,权值的初始化范围为 -0.5~0.5 之间。根据之前的研究可知 MIFS 算法与 MIFS-U 算法中的参数取值在 0.5~1 之间,算法性能最优,本文将这两个算法的参数取为 0.5。MIFS-FI 算法与 7 种算法在 14 个数据集上的分类准确率如图 4 所示,其中横坐标表示特征选择的个数,纵坐标表示分类准确率。由图 4 可知,在 14 组数据集上,本文提出的 MIFS-FI 算法相较于 7 种特征选择算法的分类准确率一直维持在较高水平。特别地,在 D_7 这个高维小样本数据集上,MIFS-FI 算法在分类器下的分类准确率超过大多数特征选择算法。

14 个数据集分类的 F_1 指数和召回率见表 4 和表 5,可见在大多数数据集上,本文所提出的 MIFS-FI 算法相较于其他 7 种方法的 F_1 指数和召回率更高。MIFS-FI 算法在 14 组数据集上 F_1 指数较 MIM 算法平均高 0.133 2,较 MIFS 算法平均高0.120 1,较 mRMR 算法平均高 0.143 4,较 CIFE 算法平均高 0.096 4,较 FJMJM 算法平均高 0.057 1,较 FIM 算法平均高 0.041 6,较 DCSF 算法平均高 0.032 1;MIFS-FI 算法在 14 组数据集上召回率较 MIM 算法平均高 0.113 4,较 MIFS 算法平均高 0.105 8,较 mRMR 算法平均高 0.098 9,较 CIFE 算法平均高 0.071 7,较 FJMJM 算法平均高 0.046 0,较 FIM 算法平均高 0.048 3,较 DCSF 算法平均高 0.031 6。



(a) 数据集 D_1 (b) 数据集 D_2 (c) 数据集 D_3

(d) 数据集 D_4 (e) 数据集 D_5 (f) 数据集 D_6



(g) 数据集 D_7　(h) 数据集 D_8　(i) 数据集 D_9

(j) 数据集 D_{10}　(k) 数据集 D_{11}　(l) 数据集 D_{12}

(m) 数据集 D_{13}　(n) 数据集 D_{14}

图 4　14 个数据集上不同特征选择方法准确率

Fig. 4　Accuracy of different feature selection methods on 14 datasets

表 4　14 个数据集上的 F_1 值

Tab. 4　F_1 score on 14 data sets

数据集	F_1 值							
	MIM	MIFS	mRMR	CIFE	FJMJM	FIM	DCSF	MIFS-FI
D_1	0.682	0.762	0.853	0.718	0.785	0.830	0.832	**0.869**
D_2	0.771	0.854	0.812	0.631	0.771	0.864	0.861	**0.878**
D_3	0.723	0.784	0.734	0.815	0.883	0.815	0.862	**0.896**
D_4	0.674	0.715	0.714	**0.862**	0.810	0.820	0.845	0.853
D_5	0.823	0.781	0.737	0.850	0.864	0.881	0.871	**0.908**
D_6	0.773	0.762	0.621	0.704	0.801	0.827	**0.897**	0.883
D_7	0.684	0.743	0.783	0.762	0.761	0.782	0.786	**0.857**
D_8	0.633	0.706	0.643	0.721	0.771	0.882	0.819	**0.872**
D_9	0.791	0.812	0.784	0.843	0.886	0.823	0.863	**0.901**
D_{10}	0.763	0.722	0.731	0.809	**0.895**	0.868	0.842	0.865
D_{11}	0.827	0.741	0.822	0.861	0.784	0.879	0.878	**0.916**
D_{12}	0.756	0.823	0.814	0.814	0.872	0.863	0.883	**0.921**
D_{13}	0.761	0.721	0.736	0.812	0.897	0.794	0.832	**0.879**
D_{14}	0.743	0.784	0.813	0.811	0.783	0.852	0.842	**0.864**

表 5 14 个数据集上的召回率

Tab. 5 Recall rate on 14 data sets

数据集	平均召回率							
	MIM	MIFS	mRMR	CIFE	FJMJM	FIM	DCSF	MIFS-FI
D_1	0.743	0.762	0.801	0.810	0.805	0.823	0.862	**0.893**
D_2	0.823	0.814	0.762	0.741	0.783	0.815	0.851	**0.881**
D_3	0.760	0.792	0.754	0.785	0.753	0.765	0.812	**0.843**
D_4	0.824	0.705	0.814	0.812	0.790	0.783	**0.875**	0.862
D_5	0.831	0.801	0.781	0.790	0.838	0.826	**0.881**	0.873
D_6	0.840	0.831	0.776	0.821	0.877	0.828	0.814	**0.906**
D_7	0.724	0.673	0.681	0.736	0.835	0.865	0.846	**0.896**
D_8	0.783	0.812	0.731	**0.875**	0.866	0.859	0.846	0.873
D_9	0.804	0.791	0.763	0.823	0.846	0.864	0.853	**0.913**
D_{10}	0.716	0.732	0.681	0.794	0.826	0.826	0.852	**0.872**
D_{11}	0.843	0.792	0.782	0.872	0.831	0.881	0.872	**0.891**
D_{12}	0.736	0.814	0.820	0.792	0.862	0.826	0.843	**0.901**
D_{13}	0.756	0.734	0.741	0.824	**0.893**	0.846	0.822	0.841
D_{14}	0.729	0.762	0.821	0.817	0.846	0.813	0.825	**0.851**

为了验证所提出算法的有效性,本文做了显著性检验,结果见表 6,大部分都是接受原假设。其中原假设为 H_0,表示算法结果与原来类别之间无显著性差异,$p \geqslant 0.05$ 表示接受原假设,$p < 0.05$ 表示拒绝原假设。

表 6 算法显著性检验结果

Tab. 6 Significance test results of the algorithm

数据集	MIM	MIFS	mRMR	CIFE	FJMJM	FIM	DCSF	MIFS-FI
D_1	0.242 2	0.611 2	0.510 3	0.614 5	0.214 4	0.252 1	0.565 8	0.743 3
D_2	0.032 6	0.312 1	0.411 3	0.311 4	0.423 0	0.424 1	0.061 1	0.665 6
D_3	0.262 1	0.584 3	0.265 1	0.261 1	0.416 0	0.224 1	0.651 6	0.756 3
D_4	0.352 3	0.422 3	0.624 1	0.266 1	0.522 3	0.688 1	0.755 1	0.836 5
D_5	0.035 4	0.531 4	0.325 1	0.452 5	0.422 6	0.654 2	0.526 5	0.623 1
D_6	0.064 4	0.023 5	0.354 1	0.354 1	0.751 1	0.225 3	0.463 2	0.893 2
D_7	0.013 5	0.034 1	0.135 5	0.245 3	0.181 3	0.741 1	0.324 1	0.741 2
D_8	0.321 4	0.210 6	0.613 0	0.423 3	0.259 1	0.286 4	0.623 4	0.655 3
D_9	0.512 7	0.342 6	0.753 3	0.821 3	0.296 3	0.542 1	0.265 4	0.462 6
D_{10}	0.154 3	0.351 5	0.632 3	0.544 2	0.547 3	0.354 8	0.421 0	0.454 2
D_{11}	0.235 6	0.562 3	0.429 3	0.396 2	0.816 3	0.582 4	0.675 4	0.843 4
D_{12}	0.861 2	0.763 1	0.561 3	0.664 1	0.823 1	0.645 7	0.263 4	0.764 2
D_{13}	0.623 4	0.732 6	0.264 3	0.001 3	0.356 0	0.442 8	0.712 3	0.561 6
D_{14}	0.626 1	0.342 4	0.545 4	0.054 2	0.423 1	0.382 5	0.655 6	0.845 3

5 结束语

本文提出了一种基于特征交互的 MIFS 算法,即 MIFS-FI 算法,并将其应用于医疗诊断数据。MIFS-FI 算法解决了 MIFS 算法中冗余项系数不确定和冗余项与相关项不可比的问题,并且重视了交互信息在冗余项中的作用,进而在实现最大相关的同时,也能兼顾最大交互和最小冗余。

实验结果来看,在分类准确率、F_1 指数和召回率三方面上,MIFS-FI 算法相比 7 种基于互信息的

特征选择方法，整体性能优于其他算法。尤其在处理高维小样本数据集时，MIFS-FI 算法的分类准确率超过大多数特征选择算法，且 F_1 指数和召回率也相对高于其它特征选择方法。

MIFS-FI 算法也存在一些缺点，当交互信息在冗余信息中的占比非常小的极端情况下，会导致冗余项非常大，冗余项的系数起不到平衡冗余项和相关项的作用，可能会使与类别相关大且冗余小的特征被剔除。由于医疗诊断数据的诊断指标之间存在较大的交互性，因而在使用 MIFS-FI 算法进行特征选择时并没有出现这种情况。若将此算法应用于其他数据集上，可能会存在上述问题。

参考文献

[1] 董梦茹. 基于模糊理论和机器学习的疾病诊断方法的研究与实现[D]. 南京：南京理工大学，2021.

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

[3] 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.

[4] 李春晓. 基于信息论的有监督特征选择算法研究[D]. 长春：吉林大学，2022.

[5] 姜文煊，段友祥，孙歧峰. 基于交互信息的混合特征选择算法[J]. 应用科学学报，2021，39(4)：545-558.

[6] 陈昊楠，金敏. 基于特征交互与权重集成的癌症分类方法[J]. 计算机应用研究，2021，38(4)：1051-1057.

[7] 顾翔元，郭继昌，李重仪，等. 基于对称不确定性和三路交互信息的特征子集选择算法[J]. 天津大学学报(自然科学与工程技术版)，2021，54(2)：214-220.

[8] BATTITI R. Using mutual information for selecting features in supervised neural net learning[J]. IEEE Transactions on neural networks，1994，5(4)：537-550.

[9] KWAK N，CHOI C H. Input feature selection for classification problems[J]. IEEE transactions on neural networks，2002，13(1)：143-159.

[10] PENG H，LONG F，DING C. Feature selection based on mutual information criteria of max-dependency，max-relevance，and min-redundancy[J]. IEEE Transactions on pattern analysis and machine intelligence，2005，27(8)：1226-1238.

[11] ESTÉVEZ 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.

[12] ZHANG P，WANG X，LI X，et al. EEG feature selection based on weighted-normalized mutual information for mental fatigue classification[C]//2016 IEEE International Instrumentation and Measurement Technology Conference Proceedings. IEEE，2016：1-6.

[13] HU L，GAO L，LI Y，et al. Feature-specific mutual information variation for multi-label feature selection[J]. Information Sciences，2022，593：449-471.

[14] LIN X，LI C，REN W，et al. A new feature selection method based on symmetrical uncertainty and interaction gain[J]. Computational Biology and Chemistry，2019，83：107149.

[15] BENNASAR M，SETCHI R，HICKS Y. Feature interaction maximization[J]. Pattern Recognition Letters，2013. 34(14)，1630-1635.

[16] SALEM O A M，LIU F，SHERIF A S，et al. Feature selection based on fuzzy joint mutual information maximization[J]. Mathematical Biosciences and Engineering，2021，18(1)：305-327.

[17] WAN J，CHEN H，YUAN Z，et al. A novel hybrid feature selection method considering feature interaction in neighborhood rough set[J]. Knowledge-Based Systems，2021，227：107167.

[18] GU X，GUO J. A feature subset selection algorithm based on equal interval division and three-way interaction information[J]. Soft Computing，2021，25：8785-8795.

(上接第 130 页)

[4] BLUM A，DWORK C，MCSHERRY F，et al. Practical privacy：The SuLQ framework[C]// Twenty-Fourth ACM Sigmod-Sigact-Sigart Symposium on Principles of Database Systems. ACM.2005：128-138.

[5] 李杨，郝志峰，温雯，等. 差分隐私保护 k-means 聚类方法研究[J]. 计算机科学，2013，40(3)：287-290.

[6] 吴伟民，焕坤. 基于差分隐私保护的 DP-DBScan 聚类算法研究[J]. 计算机工程与科学，2015，37(4)：830-834.

[7] 傅彦铭，李振铎. 基于拉普拉斯机制的差分隐私保护 k-means++聚类算法研究[J]. 信息网络安全，2019(2)：43-52.

[8] 郑孝遥，陈冬梅，刘雨晴，等. 基于差分隐私保护的谱聚类算法[J]. 计算机应用，2018，38(10)：2918-2922.

[9] 高瑜，田丰，吴振强. 基于差分隐私保护的 DPk-medoids 聚类算法[J]. 计算机技术与发展，2017，27(10)：117-120，125.

[10] DWORK C，MC SHERRY F，NISSIM K，et al. Calibrating noise to sensitivity in private data analysis[C]//Proceeding of the 39th Theory of Cryptography Conference.New York：ACM Press，2006：363-385.

[11] 孔钰婷，谭富祥，赵鑫，等. 基于差分隐私的 K-means 算法优化研究综述[J]. 计算机科学，2022，49(2)：162-173.