# 基于最大相关最小冗余联合互信息的多标签特征选择算法

张俐[1,2]，王枞[1,2]

（1. 北京邮电大学软件学院，北京 100876；2. 北京邮电大学可信分布式计算与服务教育部重点实验室，北京 100876）

**摘 要**：在过去的几十年中，特征选择已经在机器学习和人工智能领域发挥着重要作用。许多特征选择算法都存在着选择一些冗余和不相关特征的现象，这是因为它们过分夸大某些特征重要性。同时，过多的特征会减慢机器学习的速度，并导致分类过渡拟合。因此，提出新的基于前向搜索的非线性特征选择算法，该算法使用互信息和交互信息的理论，寻找与多分类标签相关的最优子集，并降低计算复杂度。在 UCI 中 9 个数据集和 4 个不同的分类器对比实验中表明，该算法均优于原始特征集和其他特征选择算法选择出的特征集。

**关键词**：特征选择；条件互信息；特征交互；特征相关；特征冗余

**中图分类号**：TP181

**文献标识码**：A

**doi:** 10.11959/j.issn.1000-436x.2018082

# Multi-label feature selection algorithm based on joint mutual information of max-relevance and min-redundancy

ZHANG Li[1,2], WANG Cong[1,2]

1. School of Software, Beijing University of Posts and Telecommunications, Beijing 100876, China
2. Key Laboratory of Trustworthy Distributed Computing and Service of Ministry of Education, Beijing University of Posts and Telecommunications, Beijing 100876, China

**Abstract:** Feature selection has played an important role in machine learning and artificial intelligence in the past decades. Many existing feature selection algorithm have chosen some redundant and irrelevant features, which is leading to overestimation of some features. Moreover, more features will significantly slow down the speed of machine learning and lead to classification over-fitting. Therefore, a new nonlinear feature selection algorithm based on forward search was proposed. The algorithm used the theory of mutual information and mutual information to find the optimal subset associated with multi-task labels and reduced the computational complexity. Compared with the experimental results of nine datasets and four different classifiers in UCI, the proposed algorithm is superior to the feature set selected by the original feature set and other feature selection algorithms.

**Key words:** feature selection, conditional mutual information, feature interaction, feature relevance, feature redundancy

## 1 引言

近年来，大数据、云计算和人工智能等技术的迅速发展，给人类社会生产和生活带来了前所未有的变化，在这之中就产生了大量的数据[1]。这些数据逐渐呈现出复杂化与高维化的趋势，同时，这些高维数据存在着大量的冗余性和无关性的特征。而这些特征增加了机器学习算法的复杂度和运行时间，同时降低了模型预测的准确性，由此带来了“维数灾难”的问题[2]。特征选择就是通过选取具有代表性的特征子集，用选择好的特征子集代替全集进行学习模型的构建与训练。简单说，特征选择可以通过降低特征维数，提升学习模型的训练速度从而达到比较好的训练效果。它通常的做法是通过挖掘

**收稿日期**：2017-09-27；**修回日期**：2018-04-18

**基金项目**：国家科技基础性工作专项基金资助项目（No.2015FY111700-6）

**Foundation Item:** The National Science and Technology Basic Work Project (No.2015FY111700-6)

特征与目标对象的相关性以及特征间冗余性的关系寻找最佳的特征子集对象。显然，特征选择算法已经成为大数据、云计算、人工智能背景下企业商务活动和经济决策的重要研究方向，从而引起了国内外众多学者的关注[3~6]。目前，常见的降维方法主要分为 2 类：特征提取和特征选择。特征提取方法主要是将已经存在的特征集转换为一种新的低维特征空间，新的特征集合是以线性或非线性的方式创建的，它代表的方法有线性判别分析[7]（LDA, linear discriminate analysis）、独立成分分析（ICA, independent component analysis）和主成分分析[8]（PCA, principal component analysis）；而特征选择方法是从原始特征集合中选择出一些最有效的特征以降低数据集维度的过程。目前，特征选择技术大量被应用到数据挖掘、机器学习、图像处理和自然语言处理等方面。特征选择算法主要分为 2 类：一种是依赖分类器模式（wrapper 方法[9]和 embedded 方法），另一种是不依赖分类器模式（filter 方法）。wrapper 方法是把各种特征子集当作测试对象，然后通过某个分类器的分类准确率作为该组特征子集性能度量指标，最后根据测试的结果得到最优的特征子集。它的主要问题为：一是计算量非常巨大；二是它采用了穷举的方式去搜索所有可能的特征组合，方法笨拙且不利于推广；三是对某个分类器过分依赖，容易出现“过拟合”现象。embedded 方法集成在学习训练过程中，较 wrapper 方法计算简单而且出现“过拟合”较少，但是寻找适应当前样本的函数模型非常困难。filter 方法依据特征与标签的相关性进行特征排序。filter 方法主要优点有 3 个：一是在特征降维方面高效和扩展性强；二是它独立于某个具体的分类器；三是信息论理论广泛应用在 filter 方法中，使 filter 方法拥有比较扎实的理论基础。例如，文献[10~17]提出的互信息、交互信息、条件互信息和联合互信息等。

本文将依赖于信息论的基本理论知识，提出一种新的基于非线性的特征选择方法——JMMC。该算法的目的就是要克服当前 filter 方法的某些局限性，例如，对某些特征的高估往往会导致其相关性和冗余性无法识别等问题。同时，本文算法借用前向搜索方法和最大最小的特征选择原则，得到更好的特征子集，它能更高效地处理高维数据集。在 UCI 中的 9 个公开数据集中进行实验，结果显示本文所提 JMMC 方法能够有效地降低数据维度，并且也能够提高不同分类器的分类准确度。

## 2 相关工作

filter 方法[18~24]是当前特征选择的一个研究热点。因此，本节将详细介绍最近几年基于 filter 方法的特征选择算法。

互信息法[6,14]主要依赖特征与目标对象的互信息值大小来排列特征的次序。当然在其中可能存在若干个冗余特征。因此，它对于高维特征集合就显得力不从心。

Kwak 等[18]提出了统一信息分布下的互信息特征选择（MIFS-U, mutual information feature selection under uniform information distribution）方法去改进 MIFS 方法，该方法依然采用特征与目标对象去衡量互信息的值，只不过它采用的是基于贪婪性的方法去搜索最有用的特征。

文献[19,20]提出了 MIFS 方法的改进版本——归一化互信息的特征选择方法（NMIFS），它用标准化的互信息取代了原来的互信息，这种做法的优点可以避免互信息值在多值时取 0 的可能。

文献[21]提出了联合互信息的特征选择方法（JMI, joint mutual information），他们认为在联合互信息中所选特征子集中特征累加和值最大就是候选特征，那么由这些候选特征构成的特征子集就是最优子集。

Peng 等[22]提出了最小冗余最大相关特征搜索（minimal redundancy and maximal relevance）算法以及标准 mRMR[23]。它们都使用基于互信息值对特征与类标签集 $C$ 的相关性以及与所得特征子集 $S$ 的冗余性进行打分，并将得分最高的特征加入 $S$ 中。mRMR 算法的优点是它将对特征子集的评价转化为对单个特征的评价，同时将当前特征子集中特征间的平均相关性来表示候选特征与当前特征子集之间的冗余性。

文献[15]提出了 FCBF（fast correlation basd filter solution），这是基于系统不确定性原理的特征选择方法。它通过候选特征与类别相关性来进行特征排除，然后采用一个近似马尔可夫毯原理对所得特征子集中冗余特征进行删减。

文献[24]提出了多标签的特征选择算法，它主要通过多标签的方式去寻找重要的特征，从而构成特征子集。

综上所述，目前，绝大多数特征选择算法关注

的重点依然是特征与标签之间的相关性和特征之间的冗余性。当然，上面介绍的特征选择方法都有它们各自的局限性。例如，MIFS-U 就是当所选特征在增加的时候，特征子集的数目也在增加，同时特征间的冗余性也同样在增加；mRMR 和 NMIFS 在每次计算中仅涉及 2 个特征间的冗余性度量，这很容易造成某些特征重要性被过分夸大。

## 3　最大联合互信息算法分析与研究

### 3.1　熵、条件熵、互信息和交互信息

**定义 1**　熵[25]是香农在 1948 年提出的，它主要用来解决信息量化度量的问题。设随机变量 $Y=\{y_1,\cdots,y_n\}$，$p(y_i)$ 为 $y_i$ 的先验概率，那么 $H(Y)$ 的熵就可以表示为

$$H(Y)=-\sum_{i=1}^{n}p(y_i)\mathrm{lb}p(y_i) \tag{1}$$

从式(1)可知，$Y$ 不确定性越大，$H(Y)$ 也就越大，那么所需要的信息量也就越大。

**定义 2**　设随机变量 $X=\{x_1,\cdots,x_n\}$，$p(x_i)$ 为 $x_i$ 的先验概率，$Y=\{y_1,\cdots,y_m\}$，$p(y_j)$ 为 $y_j$ 的先验概率，那么随机变量 $X$ 和随机变量 $Y$ 的联合可以熵表示为

$$H(X,Y)=-\sum_{i=1}^{n}\sum_{j=1}^{m}p(x_iy_j)\mathrm{lb}p(x_iy_j) \tag{2}$$

**定义 3**　设随机变量 $X=\{x_1,\cdots,x_n\}$，$p(x_i)$ 为 $x_i$ 的先验概率，$Y=\{y_1,\cdots,y_m\}$，$p(y_j)$ 为 $y_j$ 的先验概率，那么随机变量 $Y$ 下随机变量 $X$ 的条件熵可以表示为

$$H(X|Y)=-\sum_{i=1}^{n}\sum_{j=1}^{m}p(x_iy_j)\mathrm{lb}p(x_i|y_j) \tag{3}$$

**定义 4**　条件熵与联合熵之间的关系如式(4)所示。

$$H(X,Y)=H(X)+H(Y|X)=H(Y)+H(X|Y) \tag{4}$$

**定义 5**　互信息[6]是信息论里的一种信息度量工具，它表示一个随机变量 $X$ 中包含的关于另一个随机变量 $Y$ 的信息量。设变量 $X$ 和变量 $Y$，联合概率密度函数是 $p(xy)$，它们边缘概率密度函数分别是 $p(x)$、$p(y)$，那么它们的互信息 $I(X;Y)$ 可以表示为

$$I(X;Y)=\sum_{x\in X}\sum_{y\in Y}p(xy)\mathrm{lb}\frac{p(xy)}{p(x)p(y)} \tag{5}$$

同理，根据式(1)~式(5)，互信息与熵、条件熵和联合熵之间的关系可以表示为

$$I(X;Y)=H(X)-H(X|Y)=H(X)+H(Y)-H(X,Y) \tag{6}$$

**定义 6**　条件互信息。设有 3 个随机变量分别是 $X$、$Y$、$C$，它们的联合概率密度函数分别是 $p(XYC)$、$p(X|C)$ 和 $p(Y|C)$ 以及条件概率密度函数 $p(XY|C)$，假设随机变量 $C$ 是已知的，那么随机变量 $X$ 和随机变量 $Y$ 关于随机变量 $C$ 的条件互信息 $I(X;Y|C)$ 可以表示为

$$I(X;Y|C)=\sum_{x\in X}\sum_{y\in Y}\sum_{c\in C}p(XYC)\mathrm{lb}\frac{p(XY|C)}{p(X|C)p(Y|C)} \tag{7}$$

依据互信息和熵的定义，联合互信息可以表示为

$$I(X,Y;C)=I(X;C|Y)+I(Y;C) \tag{8}$$

**定义 7**　交互信息[26]是指在任何特征子集中不存在，但是它却被所有特征所共享的信息。通常，交互信息 $I(X;Y;C)$ 与联合互信息、互信息之间的关系可以表示为

$$I(X;Y;C)=I(X,Y;C)-I(X;C)-I(Y;C) \tag{9}$$

通常来说，交互信息值可以是正、负或 0。当随机变量 $X$、$Y$ 组合在一起共同提交的信息不包含它们各自提交的信息时，交互信息表示为正，而最大交互信息是指随机变量 $X$、$Y$ 组合在一起所获得的最大信息值；当随机变量 $X$、$Y$ 各自提交信息包括某些相同的一些信息时，交互信息表示为负；当随机引入某个随机变量时，并不影响随机变量 $X$（或 $Y$）和 $C$ 之间的关系时，交互信息表示为 0。

### 3.2　特征的相关性、冗余性和交互信息

特征相关性的分析是描述特征重要性的关键方法之一。在信息论中，特征的相关性根据其特点可以分为相关、冗余、无关和交互。John 等[27]将特征分为 3 类：强相关、弱相关和无关特征。而在一个最优特征集合中通常应该包括所有的强相关特征和一部分弱相关特征，不包括无关特征等。由于弱相关特征的组成相当复杂，如何从这些特征中进一步筛选出冗余特征，对特征选择算法性能提升至关重要。文献[18~24]提出了许多特征选择算法中都存在着选择了一些冗余和不相关的特征。

因此，下面将结合信息论中互信息、条件互信息和交换信息的概念，给出判断特征相关性的标准。

**定义 8** 假设数据样本集 $D=(T,F,C)$，$n$ 表示样本的数量，样本空间维数为 $m$，$T=\{t_1,\cdots,t_n\}$ 表示样本集合，$C=\{c_1,\cdots,c_k\}$ 表示标签集合，$F=\{f_1,\cdots,f_m\}$ 表示特征集合。$f_i\in F$，$f_j\in F$，其中 $f_i\neq f_j$。$S\subset F$ 表示 $S$ 是 $F$ 的子集，$S'=F-S$ 表示 $S'$ 是 $F$ 的子集。

**定义 9** 特征相关性。在 $S$ 已知的情况下，当 $f_i\notin S$、$f_j\notin S$ 时，如果存在 $I(f_i,C;S)>I(f_j,C;S)$，就表示 $f_i$ 与目标标签 $C$ 的相关性大于 $f_j$ 与目标标签 $C$ 的相关性。那么，可以进一步推导出，如果 $f_i$ 与 $f_j$ 存在依赖关系，并且 $f_i\notin S$，$S=f_j\cup S$，就可以说明特征 $f_i$ 与目标标签 $C$ 的相关性会因为 $f_j$ 的加入而提高，即 $I(f_i,C;S)>I(f_i,C)$。

**定义 10** 特征冗余性。在 $S$ 已知的情况下，如果 $f_i$ 与 $f_j$ 存在依赖关系，$f_j$ 是冗余特征，并且 $f_i\notin S$，$S=f_j\cup S$，那么就可以说明特征 $f_i$ 与目标标签 $C$ 的相关性会因为 $f_j$ 的加入而减少，即 $I(f_i,C;S)<I(f_i,C)$。同时，也可以进一步表示为 $I(F;C)\leqslant I(S;C)+I(S';C)$。

**定义 11** 特征无关性。在 $S$ 已知的情况下，如果 $f_i$ 与 $f_j$ 之间是无关的，即表示 $f_i$ 与 $f_j$ 相互独立，那么就可以说明特征 $f_i$ 与目标标签 $C$ 的相关性不会因为 $f_j$ 的加入而提高，即 $I(f_i,C;S)=I(f_i,C)$。

**定义 12** 特征间交互信息。设 $f_i\in S'$，$f_j\in S$，根据式(9)，如果 $I(F;C)\geqslant I(S;C)+I(S';C)$，就表示在 $F$ 集合中的任何一个特征的缺失都会导致降低对目标标签 $C$ 的预测能力。同时，交互信息值还可以进一步表示为当不同的特征 $f_j$ 加入 $S$ 时，特征 $f_i$ 与目标标签 $S$ 的相关性的值就可以表示为 $I(f_i,C;f_j)$，那么加入的特征的最小值就可以表示为 $\min\limits_{j=1,2,\cdots,k} I(f_i,C;f_j)$。

### 3.3 算法推导

从 3.2 节可以知道，特征之间、特征与目标标签 $C$ 之间的相关性完全可以通过 $I(S',C;S)$ 与 $I(S';C)$、$I(S';S)$ 之间的大小差值进行确定。因此，$I(S',C;S)$ 与 $I(S';C)$、$I(S';S)$ 之间的关系决定了该特征是否是相关特征、冗余特征和无关特征。

根据文献[3]和文献[11]，给出特征评价准则为

$$J(F)=D(F)-R_s(F)-C_s(F) \tag{10}$$

其中，$D(F)$ 表示特征 $F$ 与标签 $C$ 的相关性；$R_s(F)$ 表示特征 $F$ 与所选择集合 $S$ 中的特征之间的冗余性；$C_s(F)$ 表示特征 $F$ 与所选择集合 $S$ 中的特征之间的交互信息。

再根据式(8)~式(10)以及文献[21]就可以推导出最终所要的结果

$$f_{\mathrm{JMMC}}=\arg\max_{f_i\in F-S}(\min_{f_j\in S}(f_i,C;f_j)) \tag{11}$$

将式(11)右边计算式展开，设 $f_i$、$f_j$ 为特征，$C$ 为标签，$f_i\in F-S$，$f_j\in S$，$f_i\neq f_j$，根据交互信息规则[24]又有

$$I(C;f_i;f_j)=I(f_i;f_j\mid C)-I(f_i,f_j) \tag{12}$$

结合式(4)~式(12)，有

$$f_{\mathrm{JMMC}}=\arg\max_{f_i\in F-S}[I(f_i,C)-(\min_{f_j\in S}((f_i,C;f_j)-I(f_i,f_j)] \tag{13}$$

在实际中，$I(f_i;f_j\mid C)$ 求解非常困难，因此本文结合文献[11,28~30]给出一种近似求解方法。

令

$$W_{i,j}=\frac{I(f_i,f_j)}{H(f_j)}=1-\frac{H(f_j\mid f_i)}{H(f_j)} \tag{14}$$

当 $C$ 给定时，$W_{i,j}$ 会因为 $C$ 的存在而不发生任何变化。因此，可以得出

$$\frac{I(f_i,f_j)}{H(f_j)}=\frac{I(f_i,f_j\mid C)}{H(f_j\mid C)} \tag{15}$$

最后综合式(13)~式(15)可以得出

$$f_{\mathrm{JMMC}}=\arg\max_{f_i\in F-S}[I(f_i,C)-(\min_{f_j\in S}(\frac{I(f_i,f_j)}{H(f_j)}H(f_j\mid C)-I(f_i,f_j)] \tag{16}$$

### 3.4 算法描述

通过上面的分析可以知道，在进行特征选择时，除了要充分考虑特征相关性、冗余性外，还要考虑特征间与标签之间的交互信息，而在第 2 节所分析的特征排序算法中，有些算法只是考虑了特征与标签之间的相关性；有些算法在考虑特征与标签之间的相关性以及特征与特征之间的冗余性的同时，却往往忽略了特征之间还有交互信息的存在[31~36]。因此，本文提出了最大联合互信息算法（JMMC），充分考虑了特征与标签之间的相关性，更进一步考虑了某些重要特征对标签乃

至整个数据集产生的影响，并且也兼顾考虑了特征与特征之间交互信息，同时也借用了最大相关最小冗余的思想。下面，给出 JMMC 特征选择算法伪代码实现，具体如算法 1 所示。

**算法 1**　JMMC 特征选择

**输入**　原始数据集 $D$；原始特征集 $F$；类标签集合 $C$

**输出**　期望所选的特征排序集 $S$

1) 初始化

2) $S \leftarrow \phi$;

3) 计算最大互信息

4) for each　$\forall f_i \in F$ do

5) 计算每一个特征的互信息 $I(f_i;C)$，并存入 relevant_mi_set 集合中

6) $f_{\max}$ =max_sort(relevant_mi_set)

7) $F \leftarrow F \setminus f_{\max}; S \leftarrow f_{\max}$

8) 使用贪婪搜索方法寻找下一个特征

9)　repeat until $F$ 集合不为空：

10) 选择下一个特征方法：

11) 根据式(16)进行计算

12) $F \leftarrow F \setminus f_{\max}$

13) $S \leftarrow S \bigcup f_{\max}$

14) 当 $F$ 集合为空时，跳出循环

15) 输出所要的排序集合 $S$

步骤 1)~步骤 2)，初始化最优特征集 $S$；步骤 3)~步骤 7)，选择和标签类别相关性最大的特征变量，存入 $S$ 集合中；步骤 8)~步骤 14)，使用前向贪婪性搜索方法并结合式(16)，得到与标签最大相关且与其他特征两两之间最小冗余的特征 $f_{\max}$ 并加入 $S$ 中，循环结束的标志是 $F$ 特征集合为空。步骤 15)，算法就得到了最优集 $S$。

## 4　实验研究

本节将对 JMMC 算法进行有效性验证，主要从以下 2 个方面证明其有效性：1) 看 JMMC 算法是否具有特征降维效果，并能同时提高模型的准确度；2) 与其他特征选择方法相比较，JMMC 算法是否能有更好的降维效果。本文实验的研究框架具体如图 1 所示。

### 4.1　特征选择方法

JMMC 算法在设计之初，就充分考虑了特征与标签的相关性以及特征之间的冗余性和交互信息，目的是有效地识别在特征集合中是否有冗余特征和无关特征的存在。为了解决上面实验考虑的问题，本文选择 4 类具有代表性的特征选择方法作为 JMMC 算法的比较对象，它们分别是 FullSet、IG、ReliefF 和 FCBF。

1) FullSet 方法就是原始特征集合，选择它的目的就是要研究 JMMC 算法做的特征选择是否真的有效，即是否具有特征降维效果，并同时提高模型的准确度。

2) IG 全称为信息增益（information gain）[6,14]，是一种经典的特征排序算法，它主要研究候选特征与标签的相关性，也就是说候选特征与分类标签越相关，它们之间的信息增益越大，该特征越重要。最后，根据特征与标签的相关性大小顺序，进行特征排序并输出。因此，IG 特征算法是特征选择领域中检验所提算法有效性时最常见的基准算法之一。

3) ReliefF 算法[11]是一种经典的基于特征距

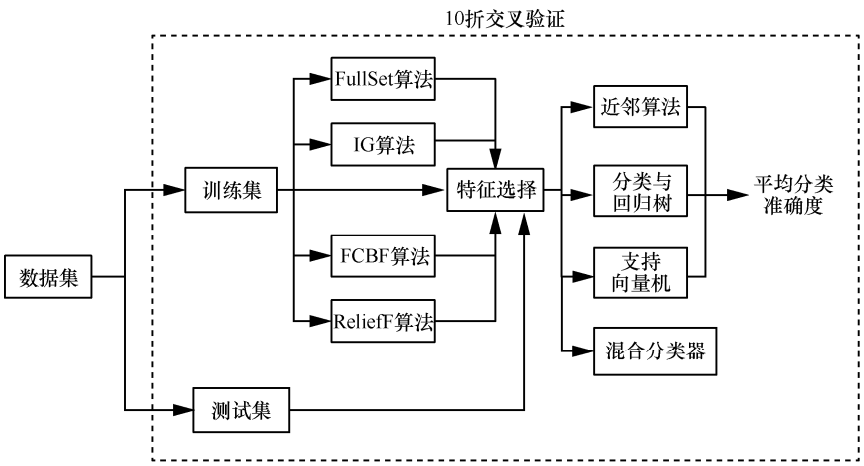


图 1　研究框架

离的排序算法。它从样本中的类内距离和类间距离来衡量特征之间的差异。一般认为好的特征应该属于同一类并且是该样本的最近的邻居，而属于不同标签的样本应该在该特征上取值尽可能不同。在本文中，ReliefF 算法中近邻数和设置为 5。

4) FCBF 算法在第 2 节已有介绍，在此不再赘述。

### 4.2 分类模型

本文的实验环境是 lenovo-ThinkPad 笔记本，处理器是 Intel(R)-Core(TM) i7-4500UCPU@ 1.80 GHz,2.4 GHz，内存是 8 GB，Windows 7 64 位操作系统，pycharm 和 Anaconda2-(64-bit)开发环境，python 的运行环境版本是 2.7.12。同时，本文采用几种频率高的分类器模型，具体如下所示。

1) 近邻（KNN, *k*-nearest neighbor）算法，指当一个样本在特征空间中有 $k$ 个最相邻的样本，而这些样本中的大多数属于某一个类别，那么该样本也属于这个类别。KNN 近邻分类器是最为经典的分类器算法。KNN 的近邻数设置为 3。

2) 分类与回归树（C4.5, classification and regression tree）算法通过 entropy 或基尼系数来选择特征进行分叉。最终将具有 $p$ 维特征的 $n$ 个样本分到 $c$ 个类别中去。在本文中，C4.5 采用基尼系数进行特征分叉。

3) 支持向量机（SVM, support vector machine），指通过升维把低维样本向高维空间做映射，使原本在低维样本空间中非线性可分的问题转化为在特征空间中线性可分的问题。SVM 分类器的参数都使用 sklearn 包的默认参数设置。

4) 混合分类器，就是将 3KNN、C4.5 和 SVM 进行混合来看它们整体的分类结果，在这里，3KNN、C4.5 和 SVM 中的权重均取 $\frac{1}{3}$。

### 4.3 实验数据集

本文选择的实验数据全部来自国际通用的 UCI 机器学习的数据集，它们分别是 heart、dermatology、movement_libra、wdbc、arrhythmia、musk、mfeat-kar、mushroom、kr-vs-kp 这 9 个数据集，详细内容如表 1 所示。样本数据分类数为 2~15 个，样本数为 270~8 124 个，样本的特征数为 14~279 维。为了保证数据更有说服性，整个实验过程采用 10 折交叉验证[37]对实验数据集进行测试和评价，最后，通过对 10 次实验求均值得到最后的实验结果。

**表 1 实验中的 UCI 数据集**

| 序号 | 数据集 | 样本数/个 | 特征数/个 | 分类/个 |
|---|---|---|---|---|
| 1 | heart | 270 | 14 | 2 |
| 2 | dermatology | 358 | 35 | 6 |
| 3 | movement_libra | 360 | 91 | 15 |
| 4 | wdbc | 569 | 31 | 2 |
| 5 | arrhythmia | 416 | 279 | 12 |
| 6 | musk | 476 | 167 | 2 |
| 7 | mfeat-kar | 2 000 | 65 | 10 |
| 8 | mushroom | 8 124 | 22 | 2 |
| 9 | kr-vs-kp | 3 195 | 37 | 2 |

### 4.4 实验结果分析

本文均采用 Accuracy 预测特征算法的优劣。同时，为了进一步说明不同算法在不同分类器和数据集的优劣，本文使用 Win/Draw/Loss 来统计并分析算法两两之间的差异。Win 表示算法 A 好于 B，Draw 表示算法 A 等于 B，Loss 表示算法 A 差于 B。

### 4.5 小样本低维数据集的分析

在小样本数据集合中，平均样本数为 390 个，平均特征数为 42.75。表 2~表 4 给出了所选择 4 种数据集上不同分类器（3KNN、C4.5 和 SVM）采用 10 折交叉验证法所获得的平均分类准确率和特征数。表 5 给出了所选择 4 种数据集上混合分类器（由 3KNN、C4.5 和 SVM 组成）采用 10 折交叉验证法所获得的平均分类准确率和特征数。图 2~图 5 给出了这 4 种数据集合的显示混合分类器效果，其中，横坐标表示依次递增的所选特征子集数目，纵坐标表示平均分类率准确率。根据表 2~表 5 以及图 2~图 5 所示的实验结果，JMMC 算法使用 3KNN 分类器在 heart 数据集、dermatology 数据集、movement_libra 数据集和 wdbc 数据集上比 FullSet 要高出 4.074%、0.833%、0.666%和 1.391%；并且，JMMC 算法使用 C4.5 分类器在 heart 数据集、dermatology 数据集、movement_libra 数据集和 wdbc 数据集上比 FullSet 要高出 1.852%、0.833%、0.666%和 1.391%；JMMC 算法使用 SVM 分类器在 heart 数据集、dermatology 数据集、movement_libra 数据集和 wdbc 数据集上比 FullSet 要高出 1.481%、0.294%、

**表 2　　基于 3KNN 分类器的所选特征集平均准确率**

| 序号 | FullSet 算法 | JMMC 算法 | | IG 算法 | | FCBF 算法 | | ReliefF 算法 | |
|---|---|---|---|---|---|---|---|---|---|
| | 准确度 | 特征 | 准确度 | 特征 | 准确度 | 特征 | 准确度 | 特征 | 准确度 |
| 1 | 67.037% | 8 | 71.111% | 8 | 66.296% | 8 | 66.296% | 6 | 66.296% |
| 2 | 96.013% | 31 | 96.846% | 28 | 91.027% | 34 | 90.455% | 24 | 97.171% |
| 3 | 80% | 87 | 80.666% | 85 | 79.222% | 85 | 79.222% | 61 | 79.555% |
| 4 | 92.633% | 16 | 94.024% | 26 | 92.633% | 26 | 92.633% | 7 | 92.633% |
| 平均值 | 83.921% | 35.5 | 85.662% | 36.75 | 82.294% | 38.25 | 82.151% | 24.5 | 83.914% |

**表 3　　基于 C4.5 分类器的所选特征集平均准确率**

| 序号 | FullSet 算法 | JMMC 算法 | | IG 算法 | | FCBF 算法 | | ReliefF 算法 | |
|---|---|---|---|---|---|---|---|---|---|
| | 准确度 | 特征 | 准确度 | 特征 | 准确度 | 特征 | 准确度 | 特征 | 准确度 |
| 1 | 76.296% | 8 | 78.148% | 7 | 76.296% | 12 | 74.814% | 11 | 75.925% |
| 2 | 94.735% | 26 | 96.124% | 25 | 94.146% | 32 | 93.297% | 29 | 95.029% |
| 3 | 68.444% | 74 | 67.777% | 53 | 66.666% | 67 | 67.333% | 27 | 68.444% |
| 4 | 95.099% | 6 | 95.437% | 16 | 94.384% | 16 | 94.739% | 23 | 94.212% |
| 平均值 | 83.643% | 28.5 | 84.371% | 25.25 | 82.873% | 31.75 | 82.545% | 22.5 | 83.402% |

**表 4　　基于 SVM 分类器的所选特征集平均准确率**

| 序号 | FullSet 算法 | JMMC 算法 | | IG 算法 | | FCBF 算法 | | ReliefF 算法 | |
|---|---|---|---|---|---|---|---|---|---|
| | 准确度 | 特征 | 准确度 | 特征 | 准确度 | 特征 | 准确度 | 特征 | 准确度 |
| 1 | 70.37% | 8 | 71.851% | 5 | 59.259% | 1 | 57.407% | 2 | 55.925% |
| 2 | 97.434% | 25 | 97.743% | 24 | 93.169% | 34 | 92.645% | 31 | 97.728% |
| 3 | 52.666% | 61 | 53.555% | 89 | 51.111% | 89 | 51.111% | 84 | 53.555% |
| 4 | 88.973% | 10 | 94.569% | 5 | 89.111% | 5 | 89.111% | 1 | 69.197% |
| 平均值 | 77.361% | 26 | 79.429% | 30.75 | 73.162% | 32.25 | 72.568% | 29.5 | 69.101% |

**表 5　　基于混合分类器和不同算法的平均准确率**

| 序号 | FullSet 算法 | JMMC 算法 | | IG 算法 | | FCBF 算法 | | ReliefF 算法 | |
|---|---|---|---|---|---|---|---|---|---|
| | 准确度 | 特征 | 准确度 | 特征 | 准确度 | 特征 | 准确度 | 特征 | 准确度 |
| 1 | 68.765 % | 4 | 79.51 % | 6 | 66.173 % | 5 | 71.111% | 6 | 66.42 % |
| 2 | 95.722% | 21 | 96.952 % | 34 | 92.416% | 22 | 96.692% | 23 | 93.522% |
| 3 | 65.888% | 81 | 66.777% | 85 | 65.148% | 87 | 64.962% | 82 | 66.037% |
| 4 | 87.301% | 7 | 93.926 % | 5 | 91.107 % | 8 | 90.288% | 5 | 91.284% |
| 平均值 | 79.419% | 28.25 | 84.291% | 32.5 | 78.711% | 30.5 | 80.763% | 29 | 79.316% |

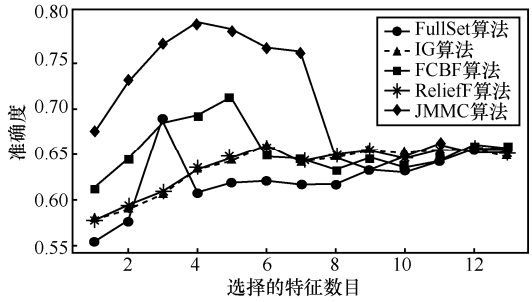


图 2 heart 数据集不同算法的平均正确率

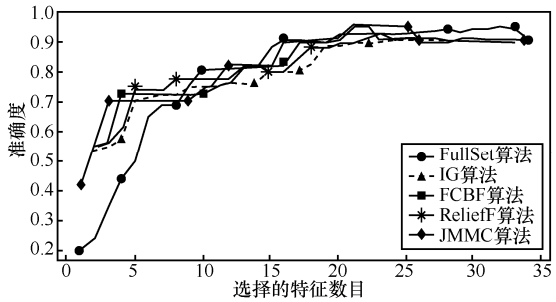


图 3 dermatology 数据集不同算法的平均正确率

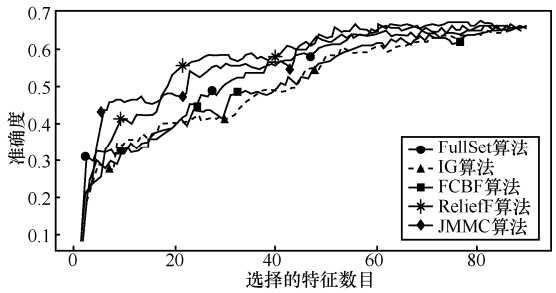


图 4 movement_libra 数据集不同算法的平均正确率

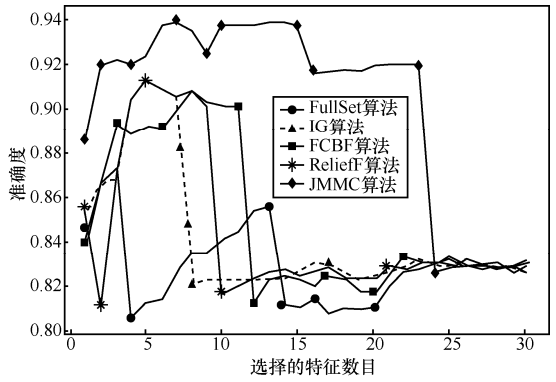


图 5 wdbc 数据集不同算法的平均正确率

0.889%和 5.596%；JMMC 算法使用混合分类器在 heart 数据集、dermatology 数据集、movement_libra 数据集和 wdbc 数据集上比 FullSet 要高出 10.745%、1.23%、0.889%和 6.625%。通过对以上 4 种数据集在不同分类器以及它们的混合分类器来看，JMMC 算法均起到了降低数据冗余度、提高分类准确度的效果，同时，从表 2~表 5 可以看出，由于分类器从弱变强，分类的效果也会变好，并且所需要的平均特征数也会相应减少。现在，再来对比 JMMC 算法和其他特征排序算法，从表 2~表 5 可以看出，JMMC 算法在绝大多数情况下均优于其他所选择的特征排序算法，具体描述可以从表 6 看出。

综上可得，在小样本集中，JMMC 算法在特征的选择上由于更多地考虑了特征间冗余性与交互信息，所以说分类的准确率更好一些。

### 4.6 大样本高维数据集的分析

在大样本数据集合中，平均样本数是 2 843 个，平均特征数为 114。表 7~表 9 给出了所选择 5 种数据集上不同分类器（3KNN、C4.5 和 SVM）采用 10 折交叉验证法所获得的平均分类准确率和特征数。表 10 给出了所选择 5 种数据集上混合分类器（由 3KNN、C4.5 和 SVM 组成）采用 10 折交叉验证法所获得的平均分类准确率和特征数。图 6~图 10 给出了这 5 种数据集合的显示混合分类器效果，其中横坐标表示依次递增的所选特征子集数目，纵坐标表示平均分类率准确率。根据表 7~表 10 以及图 6~图 10 所示的实验结果，JMMC 算法使用 3KNN 分类器在 arrhythmia 数据集、musk 数据集、mfeat-kar 数据集、mushroom 数据集和 kr-vs-kp 数据集上比 FullSet 要高出 0.933%、1.549%、0、0 和 3.538%；并且，JMMC 算法使用 C4.5 分类器在 arrhythmia 数据集、musk 数据集、mushroom 数据集和 kr-vs-kp 数据集上比 FullSet 要高出 5.456%、0.195%、0 和 0；JMMC 算法使用 SVM 分类器在 arrhythmia 数据集、musk 数据集、

**表 6 JMMC 算法与其他基于特征排序算法的 Win/Draw/Loss 分析**

| 分类器 | JMMC 与 FCBF | | JMMC 与 IG | | JMMC 与 ReliefF | |
|---|---|---|---|---|---|---|
| | 特征数 | 准确度 | 特征数 | 准确度 | 特征数 | 准确度 |
| 3KNN | 2/1/1 | 4/0/0 | 1/1/2 | 4/0/0 | 0/0/4 | 3/0/1 |
| C4.5 | 3/0/1 | 4/0/0 | 1/0/3 | 4/0/0 | 3/0/1 | 4/0/0 |
| SVM | 1/0/3 | 4/0/0 | 1/0/3 | 4/0/0 | 2/0/2 | 4/0/0 |
| 混合 | 4/0/0 | 4/0/0 | 3/0/1 | 4/0/0 | 3/0/1 | 4/0/0 |

**表 7　　基于 3KNN 分类器的所选特征集平均准确率**

| 序号 | FullSet 算法 | JMMC 算法 | | IG 算法 | | FCBF 算法 | | ReliefF 算法 | |
|---|---|---|---|---|---|---|---|---|---|
| | 准确度 | 特征 | 准确度 | 特征 | 准确度 | 特征 | 准确度 | 特征 | 准确度 |
| 5 | 66.345% | 133 | 67.278% | 155 | 67.278% | 183 | 65.601% | 40 | 65.06% |
| 6 | 77.967% | 128 | 79.516% | 153 | 79.516% | 118 | 78.203% | 83 | 78.407% |
| 7 | 97.7% | 63 | 97.7% | 64 | 97.7% | 64 | 97.7% | 64 | 97.7% |
| 8 | 100% | 6 | 100% | 11 | 100% | 17 | 100% | 19 | 98.425% |
| 9 | 92.676% | 23 | 96.214% | 36 | 96.214% | 16 | 92.614% | 36 | 95.776% |
| 平均值 | 86.937% | 70.6 | 88.155% | 83.8 | 88.141% | 79.6 | 86.823% | 48.4 | 87.073% |

**表 8　　基于 C4.5 分类器的所选特征集平均准确率**

| 序号 | FullSet 算法 | JMMC 算法 | | IG 算法 | | FCBF 算法 | | ReliefF 算法 | |
|---|---|---|---|---|---|---|---|---|---|
| | 准确度 | 特征 | 准确度 | 特征 | 准确度 | 特征 | 准确度 | 特征 | 准确度 |
| 5 | 65.846% | 168 | 71.302% | 228 | 67.177% | 172 | 67.442% | 242 | 67.229% |
| 6 | 76.926% | 152 | 77.121% | 158 | 75.694% | 162 | 75.06% | 163 | 76.31% |
| 7 | 84.849% | 17 | 84.049% | 14 | 84.299% | 11 | 85.049% | 10 | 84.8% |
| 8 | 100% | 6 | 100% | 11 | 100% | 17 | 98.425% | 19 | 99.852% |
| 9 | 100% | 23 | 100% | 36 | 100% | 16 | 100% | 36 | 100% |
| 平均值 | 85.524% | 73.2 | 86.494% | 89.4 | 85.434% | 75.6 | 85.195% | 94 | 85.638% |

**表 9　　基于 SVM 分类器的所选特征集平均准确率**

| 序号 | FullSet 算法 | JMMC 算法 | | IG 算法 | | FCBF 算法 | | ReliefF 算法 | |
|---|---|---|---|---|---|---|---|---|---|
| | 准确度 | 特征 | 准确度 | 特征 | 准确度 | 特征 | 准确度 | 特征 | 准确度 |
| 5 | 56.946% | 68 | 59.012% | 3 | 56.946% | 3 | 56.946% | 5 | 56.946% |
| 6 | 60.728% | 6 | 63.517% | 2 | 58.644% | 2 | 58.644% | 2 | 59.247% |
| 7 | 94.65% | 64 | 94.65% | 64 | 94.65% | 64 | 94.65% | 64 | 94.65% |
| 8 | 97.047% | 1 | 98.523% | 1 | 98.523% | 1 | 98.523% | 18 | 97.023% |
| 9 | 100% | 23 | 100% | 36 | 100% | 16 | 100% | 36 | 100% |
| 平均值 | 81.874% | 32.4 | 83.14% | 21.2 | 81.752% | 17.2 | 81.752% | 25 | 81.573% |

**表 10　　基于不同分类器和不同算法的平均准确率**

| 序号 | FullSet 算法 | JMMC 算法 | | IG 算法 | | FCBF 算法 | | ReliefF 算法 | |
|---|---|---|---|---|---|---|---|---|---|
| | 准确度 | 特征 | 准确度 | 特征 | 准确度 | 特征 | 准确度 | 特征 | 准确度 |
| 5 | 62.853% | 171 | 63.935% | 189 | 62.983% | 237 | 62.904% | 255 | 63.364% |
| 6 | 70.198% | 158 | 70.484% | 166 | 69.913% | 162 | 69.991% | 165 | 70.129% |
| 7 | 91.416% | 63 | 91.549% | 61 | 91.333% | 64 | 91.466% | 63 | 91.383% |
| 8 | 98.15% | 6 | 98.757% | 11 | 98.014% | 20 | 96.806% | 19 | 97.839% |
| 9 | 97.558% | 23 | 98.738% | 36 | 97.538% | 36 | 97.485% | 36 | 97.444% |
| 平均值 | 84.035% | 84.2 | 84.693% | 92.6 | 83.957% | 103.8 | 83.73% | 107.6 | 84.032% |

**表 11　　JMMC 算法与其他基于特征排序算法的 Win/Draw/Loss 分析**

| 分类器 | JMMC 与 FCBF | | JMMC 与 IG | | JMMC 与 ReliefF | |
|---|---|---|---|---|---|---|
| | 特征数 | 准确度 | 特征数 | 准确度 | 特征数 | 准确度 |
| 3KNN | 3/0/2 | 4/1/0 | 5/0/0 | 3/2/0 | 3/0/2 | 3/1/1 |
| C4.5 | 3/0/2 | 3/1/1 | 4/0/1 | 2/2/1 | 4/0/1 | 3/1/1 |
| SVM | 0/2/3 | 2/3/0 | 1/2/2 | 2/3/0 | 2/1/2 | 3/2/0 |
| 混合 | 5/0/0 | 5/0/0 | 4/0/1 | 5/0/0 | 4/1/0 | 5/0/0 |

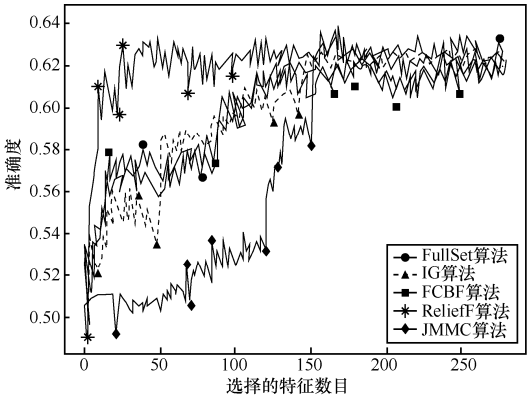


图 6　arrhythmia 数据集不同算法的正确率

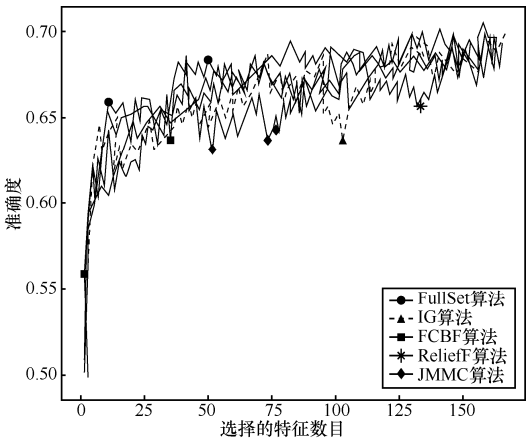


图 7　musk 数据集不同算法的平均正确率

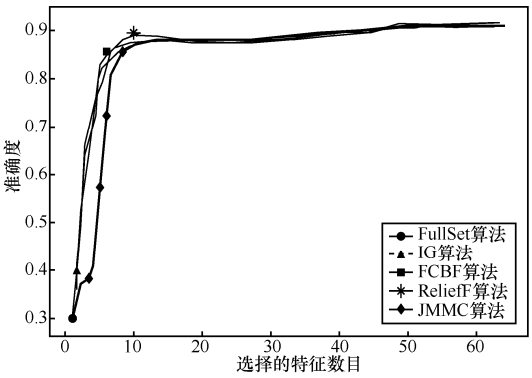


图 8　mfeat-kar 数据集不同算法的平均正确率

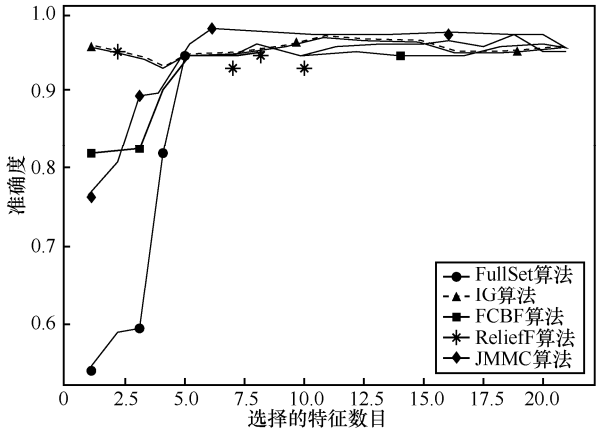


图 9　mushroom 数据集不同算法的平均正确率

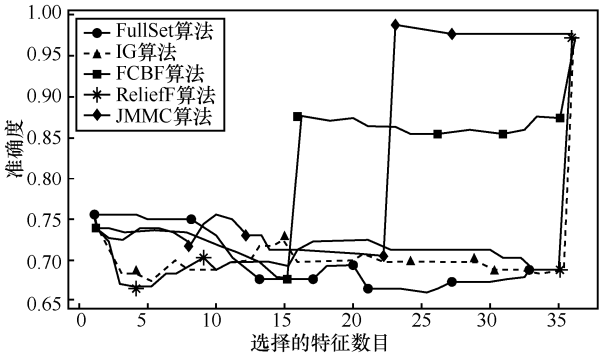


图 10　kr-vs-kp 数据集不同算法的平均正确率

mfeat-kar 数据集、mushroom 数据集和 kr-vs-kp 数据集上比 FullSet 要高出 2.066%、2.789%、0、1.476%和 0；JMMC 算法使用混合分类器在 arrhythmia 数据集、musk 数据集、mfeat-kar 数据集、mushroom 数据集和 kr-vs-kp 数据集上比 FullSet 要高出 1.082%、0.286%、0.133%、0.607%和 1.18%。从以上的数据分析可以看出，JMMC 算法只是在使用 C4.5 分类器时在 mfeat-kar 数据集上略低于 FullSet，其他均优于或等同于 FullSet。同时，通过对以上 5 种数据集在不同分类器以及它们的混合分类器来看，JMMC 算法起到了降低

数据冗余度、提高分类准确度的效果，同时，从表 8~表 10 可以看出，由于分类器从弱变强，分类的效果也会变好，并且，所需要的平均特征数也会相应减少，其中在基于 SVM 的分类器表现得尤为明显，特征数平均只需要 32.4 个。现在，再来对比 JMMC 算法和其他特征排序算法，从表 7~表 9 可以看出，JMMC 算法大部分均优于其他所选择的特征排序算法，可能由于分类器的原因造成在个别数据集上和不同分类器上存在 JMMC 算法略低于其他特征排序算法，这些具体情况都可以从表 11 看出。现在，可以得出在大样本集中，JMMC 算法在特征的选择上由于充分地考虑了特征间冗余性与交互信息，分类准确率相比其他算法要更好一些。

## 5　结束语

特征选择算法主要是尽可能寻找较小的特征子集，从而获得较高分类预测准确率。本文通过引入条件互信息和交互信息，并依赖最大最小原则建立新的特征排序算法（JMMC 算法）。它不仅考虑特征的相关性、冗余性和无关性，也充分考虑了特征间与目标标签之间的交互信息。首先，依据信息论理论，重新定义了相关性、冗余性、无关性和交互信息。其次，给出了 JMMC 算法的推导和实现过程。在 UCI 公开的 9 个样本集和 4 种不同的分类器（3KNN、C4.5、SVM 和混合分类器）中，JMMC 算法在绝大多数情况下，平均分类准确率均优于其他特征排序算法。

综上可知，JMMC 算法不仅能有效识别出相关特征、冗余特征和无关特征，而且也能识别出特征间产生交互信息的那些特征。因此，JMMC 算法可以有效地提高分类的准确度，并且降低特征的维数。下一步的工作将是对交互信息、条件互信息等理论进行更进一步的研究，以便能提出更加有效的特征排序算法来优化选择出的特征子集并进一步提高它们分类的准确率。

## 参考文献：


[1] GEORGE G, HAAS M R, PENTLAND A. Big data and management[J]. Academy of Management Journal,2014,57(2):321-326.

[2] XIE J Y, XIE W X. Several selection algorithms based on the discernibility of a feature subset and support vector machines[J].Chinese Journal of Computers, 2014,37(8):1704-1718.

[3] BROWN G, POCOCK A, ZHAO M J, et al. Conditional likelihood maximisation-a unifying framework for information theoretic feature selection[J].Journal of Machine Learning Research, 2012,13:27-66.

[4] CHENG H G, QIN Z, FENG C, et al. Conditional mutual information based feature selection analysing for synergy and redundancy[J]. Electronics and Telecommunications Research Institute, 2011(33): 210-218.

[5] CHANDRASHEKAR G, SAHIN F .A survey on feature selection methods[J].Computers and Electrical Engineering, 2014(40):16-28.

[6] ZHANG Z H, LI S N, LI Z G, et al. Multi-label feature selection algorithm based on information entropy[J].Journal of Computer Research and Development, 2013,50(6):1177-1184.

[7] YU H, YANG J.A direct LDA algorithm for high-dimensional data with application to face recognition[J].Pattern Recognition, 2001(34): 2067-2070.

[8] BAJWA I S, NAWEED M S, ASIF M N, et al. Feature based image classification by using principal component analysis[J].ICGST International Journal on Graphics Vision and Image Processing, 2009(9): 11-17.

[9] MALDONADO S, WEBER R. A wrapper method for feature selection using support vector machine[J].Information Science, 2009, 179(13): 2208-2217.

[10] PENG C. Distributed *K*-Means clustering algorithm based on Fisher discriminant ratio[J].Journal of Jiangsu University, 2014, 35(4): 422-427.

[11] ZHANG Y S, YANG A, XIONG C, et al. Feature selection using data envelopment analysis[J].Knowledge-Based Systems,2014(64):70-80.

[12] YU L, LIU H. Feature selection for high-dimensional data: a fast correlation-based filter solution[C]//The 20th International Conferences on machine learning. 2003: 856-863.

[13] HUANG D, CHOW T W S. Effective feature selection scheme using mutual information[J]. Neurocomputing,2005 (63):325-343.

[14] LIU H W, SUN J G, LIU L, et al. Feature selection with dynamic mutual information[J]. IEEE Transactions on Neural Networks, 2009, 20(2): 189-201.

[15] DUAN H X, ZHANG Q Y, ZHANG M.FCBF algorithm based on normalized mutual information for feature selection[J].Journal Huazhong University of Science & Technology(Natural Science Edition), 2017,45(1):52-56.

[16] SUN G L, SONG Z C, LIU J L, et al. Feature selection method based on maximum information coefficient and approximate markov blanket[J]. Acta Automatica Sinica, 2017,43(5):795-805.

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

[18] KWAK N, CHOI C H. Input feature selection for classification problems[J]. IEEE Transactions on Neural Networks, 2002(13):143-159.

[19] ESTÉVEZ P A, TESMER M, PEREZ C A, et al. Normalized mutual information feature selection[J].IEEE Transaction on Neural Networks, 2009(20):189-201.

[20] HOQUE N, BHATTACHARYYA D K, KALITA J K.MIFS-ND:a mutual information-based feature selection method[J]. Expert Systems with Applications,2014,41(14):6371-6385.

[21] HOWARD H Y, JOHN M. Feature selection based on joint mutual information[C]//Advances in Intelligent Data Analysis (AIDA), Computational Intelligence Methods and Applications (CIMA), International Computer Science Conventions Rochester New York. 1999: 1-8.

[22] PENG H, LONG F, DING C, Feature selection based on mutual in-

formation: criteria of max-dependency, max-relevance, and min-redundancy[C]//IEEE Transaction on Pattern Analysis & Machine Intelligence. 2005, 27 (8): 1226-1238

[23] VINH L T, THANG N D, LEE Y K. An improved maximum relevance and minimum redundancy feature selection algorithm based on normalized mutual information[C]//Tenth International Symposium on Applications and the Internet. 2010: 395-398.

[24] LEE J, KIM D W. Mutual information-based multi-label feature selection using interaction information[J].Expert Systems with Applications, 2015(42): 2013-2025.

[25] COVER T, THOMAS J.Elements of theory[M]. New York: John Wiley & Sons, 2002.

[26] JAKULIN A. Attribute interactions in machine learning (Master thesis)[M]//Lecture Notes in Computer Science. 2003.

[27] JOHN G H, KOHAVI R, PFLEGER K. Irrelevant features and the subset selection problem[C]//The Eleventh International Conference on Machine Learning, 1994: 121-129.

[28] BENNASAR M, HICKS Y, SETCHI R. Feature selection using Joint mutual information maximisation[J]. Expert System Application, 2015(42): 8520-8532.

[29] ZHANG Y S, ZHANG Z G. Feature subset selection with cumulate conditional mutual information minimization[J]. Expert Systems with Applications, 2012,39(5):6078-6088.

[30] YU L, LIU H. Efficient feature selection via analysis of relevance and redundancy[J]. Journal of Machine Learning Research, 2004, 5(12): 1205-1224.

[31] TAPIA E, BULACIO P, ANGELONE L F. Sparse and stable gene selection with consensus SVM-RFE[J]. Pattern Recognition Letters,2012,33(2):164-172.

[32] UNLER A, MURAT A, CHINNAM R B. mr2PSO: a maximum relevance minimum redundancy feature selection method based on swarm intelligence for support vector machine classification[J]. Information Sciences, 2011(20): 4625-4641.

[33] CHE J X, YANG Y L LI L, et al. Maximum relevance minimum common redundancy feature selection for nonlinear data[J]. Information Sciences, 2017(5):68-89.

[34] CHAKRABORTY R, PAL N R. Feature selection using a neural framework with controlled redundancy[J].IEEE Transactions on Neural Networks and Learning Systems, 2015,26 (1) :35-50.

[35] AKADI A E, OUARDIGHI A, ABOURAJDINE D.A powerful feature selection approach based on mutual information[J]. International Journal of Computer Science and Network Security, 2008(8):116-211.

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

[37] NIU X T. Support vector extracted algorithm based on KNN and 10 fold cross-validation method[J].Journal of Huazhong Normal University, 2014,48(3):335-338.

[作者简介]

**张俐**（1977-），男，陕西汉中人，北京邮电大学博士生，主要研究方向为机器学习、特征工程、医疗健康数据分析挖掘。

**王枞**（1958-），女，北京人，博士，北京邮电大学教授、博士生导师，主要研究方向为智能信息处理、网络信息安全、可信计算与服务、医疗健康数据分析挖掘。