Advances in Applied Mathematics 应用数学进展, 2022, 11(5), 3018-3027
Published Online May 2022 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2022.115321

Hans汉斯

基于高维复杂数据的变量选择方法研究

郭艾堃

河北工业大学理学院，天津

收稿日期：2022年4月25日；录用日期：2022年5月19日；发布日期：2022年5月30日

摘　要

针对目前大多数基于信息论的线性累加特征选择算法的缺点和不足，将类别变量的信息，引入到对待选特征与已选特征子集的冗余性度量之中，考虑到特征与类别变量之间的对称不确定度，提出了一种新的以信息论为基础的，基于最大相关最小冗余原则的过滤式特征选择方法，并在11个公开的标准数据集上进行了验证，通过与6种其他基于信息论的特征选择方法的结果进行对比，验证了所提出的算法的有效性。

关键词

高维复杂数据，特征选择，互信息，最大相关最小冗余，条件互信息

Research on Variable Selection Method Based on High-Dimensional Complex Data

Aikun Guo

School of Sciences, Hebei University of Technology, Tianjin

Received: Apr. 25th, 2022; accepted: May 19th, 2022; published: May 30th, 2022

Abstract

In view of the shortcomings and deficiencies of most current linear accumulation feature selection algorithms based on information theory, the information of categorical variables is introduced into the redundancy measure of the feature to be selected and the selected feature subset, taking into account the relationship between the feature and the categorical variable, a new filtering feature selection method based on information theory and the principle of maximum correlation and minimum redundancy is proposed, and verified on 11 public standard datasets. The results of six other information theory-based feature selection methods are compared to verify the effectiveness of the proposed algorithm.

文章引用: 郭艾堃. 基于高维复杂数据的变量选择方法研究[J]. 应用数学进展, 2022, 11(5): 3018-3027.
DOI: 10.12677/aam.2022.115321

Keywords

High-Dimensional Complex Data, Feature Selection, Mutual Information, Maximum Correlation and Minimum Redundancy, Conditional Mutual Information

1. 引言

随着计算机技术与网络的发展，数据的采集、存储的设备与方法日益普及和不断更新，我们所需要面对的数据集从以前的简单特征数据集向具有数据存储量巨大、数据结构多样、数据属性维数高、数据相关性强、数据特征时变等特点的大规模复杂数据集发展。在现代数据分析过程中，第一步也是必不可少的步骤就是对数据进行降维。降维主要有两种方法：特征提取和特征选择。与特征提取相比，特征选择可以保留原始特征，使结果的解释具有现实意义。所以在本文的研究中，我们关注的是特征选择而不是特征提取。

按照特征选择过程与分类器之间的关系，特征选择方法一般被分为嵌入式特征选择方法(Embedded) [1] [2]，包装式特征选择方法(Wrapper) [3]和过滤式特征选择方法(Filter) [4] [5] [6]三大类。嵌入式(Embedded)特征选择方法在选择特征的过程中同时完成了模型的学习。包装式(Wrapper)特征选择方法是使用分类算法对特征选择得到的特征子集进行评价，通过使用特定分类器的预测精度作为所选特征子集质量的度量来搜索特征空间并测试所有可能的特征组合子集。过滤式(Filter)特征选择方法是独立于分类器的，不与特定的分类器相关联，它根据一些标准来评估特征的预测能力。并且基于过滤式(Filter)的特征选择方法计算量相对小，通用性强，简单高效的特点，同时也可以作为数据挖掘中的预处理步骤，与其他各种方法结合，具有良好的性质。因此本文将研究重点放在基于互信息的过滤式特征选择方法。

2. 信息论的基本概念

2.1. 熵原理及信息熵

本文算法是基于信息论的，因此本节首先介绍一些信息论的基本概念[7]。

熵概念的建立将不确定性和信息量结合到了一起，事件发生的概率越低，所蕴含的信息量就越大。

随机变量的熵是对其不确定性的度量，以及描述随机变量所需的平均信息量。对于离散的随机变量 $X=\{x_1,x_2,\cdots,x_N\}$ 和随机变量 $Y=\{y_1,y_2,\cdots,y_M\}$，信息熵定义为：

$$H(X)=-\sum_{i=1}^{N}p(x_i)\log\left(p(x_i)\right)$$

2.2. 互信息

互信息(MI)是信息论的另一个重要概念。它是新引入的变量所提供的信息量，也可以看作是两个变量共享的信息量。MI 用于量化两个变量之间的相互依赖关系。当两个变量相互独立时，MI 为零，并且随着一个变量对另一个变量的依赖程度的增加而增加。变量 Y 和变量 X 之间的互信息记为 $I(Y;X)$，可以表述为：

$$I(X;Y)=\int p(x,y)\log\frac{p(x,y)}{p(x)p(y)}\mathrm{d}x\mathrm{d}y$$

很难准确地找到概率密度函数($p(x)$，$p(y)$，$p(x,y)$)。因此，通常首先对连续变量进行离散化，然后用离散定义计算熵和互信息。

互信息可以被条件化，条件互信息 $I(Y;X_i\mid X_j)$ 量化了选择 X_j 时 X_i 提供的新判别信息，定义如下：

$$I(Y;X_iX_j)=\sum_{x_j\in X_j}p(x_j)\sum_{x_i\in X_i}\sum_{y\in Y}p(y,x_ix_j)\log\frac{p(y,x_ix_j)}{p(yx_j)p(x_ix_j)}$$

联合互信息(JMI)量化了 Y 和联合变量($X_i,\cdots,X_k$)共享的信息，定义如下：

$$I(Y;X_i,\cdots,X_k)=H(Y)-H(Y\mid X_i,\cdots,X_k)$$

将三个变量之间共享的信息量定义为交互信息，其定义如下：

$$I(Y;X_i;X_j)=I(Y;X_i)+I(Y;X_j)-I(Y;X_i,X_j)$$

$I(Y;X_i;X_j)$ 的值可以是正数、负数或零，也可以看作是两个特征共享的判别信息。如果 $I(Y;X_i;X_j)$ 为正，则表示 X_i，X_j 有冗余。如果 $I(Y;X_i;X_j)$ 为负数，则表示 X_i，X_j 是互补的。当 $I(Y;X_i;X_j)$ 为零时，表示 X_i 和 X_j 相互独立。所以条件互信息 $I(Y;X_i\mid X_j)$ 可以重写为 $I(Y;X_i\mid X_j)=I(Y;X_i)-I(Y;X_i;X_j)$。这意味着候选特征提供的新信息越多，它对应的冗余信息就越少。

3. 相关工作

基于信息论的特征选择算法主要可以分为两类，一类是基于线性累积求和思想的特征选择方法，其主要思想是通过求累积和的方式来判断特征与类别的相关性大小，从而选出符合一定标准的特征，大多数方法都是基于此种思想。还有一类是基于非线性计算的特征选择方法。原则是认为通过求和的方式会过高的估计公共冗余信息所占的比例，希望通过非线性的方式来代替求和，以降低对公共冗余信息估计的误差。

布朗[8]等人认真分析和研究了基于互信息的特征选择方法，并提出了统一的特征选择框架。我们可以看到，现有的大多数基于信息论的方法都可以从这个框架中衍生出来。实际上，这个框架是基于线性累积求和思想的线性组合。定义如下：

$$J(X_i)=I(Y;X_i)-\beta\sum_{X_j\in S}I(X_i;X_j)+\gamma\sum_{X_j\in S}I(X_i;X_j\mid Y)$$

在这个公式中，它涉及到三种信息：相关性、冗余性和互补性。当 β 和 γ 取不同的值时，可以得到不同的特征选择方法。例如，当 $\gamma=0$ 时，可以得到 MIFS 方法，当时，可以得到 JMI 方法，当 $\gamma=0$ 且 $\gamma=\frac{1}{|S|}$ 时，可以得到 mRMR，而当 $\gamma=\beta=0$ 时，能够得到 MIM。相关性和冗余是两个矛盾的方面，高相关性通常意味着高冗余，如何平衡这两个方面仍然是特征选择方法的一个悬而未决的问题。

在此基础上，Wang [9]等人又提出(最大相关性和最大独立性，MRI)特征选择算法，提出的算法独立的分类信息(独立分类信息，ICI)，即 $I(Y;X_i,X_j)=I(X_i;Y\mid X_j)+I(X_j;Y\mid X_i)$，其筛选标准为

$$J_{\mathrm{MRI}}(X_i)=I(X_i;Y)+\sum_{X_j\in S}I(X_i;Y\mid X_j)+I(X_j;Y\mid X_i)$$

基于非线性计算的特征选择方法有以下几种：条件互信息最大化(CMIM) [10]采用“最大化最小值”

的标准。通过选择最小条件互信息最大的特征，CMIM 确保所选特征既具有单独的信息量，又相互弱依赖。CMIM 定义如下：

$$J_{\mathrm{CMIM}}\left(X_i\right)=\arg\max_{X_i\in F-S}\left(\min_{X_j\in S}\left(I\left(Y;X_i\mid X_j\right)\right)\right)$$

李[11]等人认为交互信息是优于条件互信息，并提出联合互信息最大化算法(JMIM)基于最大和最小标准。这算法的评价标准如下：

$$J_{\mathrm{JMIM}}\left(X_i\right)=\arg\max_{X_i\in F-S}\left(\min_{X_j\in S}\left(I\left(Y;X_i,X_j\right)\right)\right)$$

Zeng 等人[12]提出了一种基于动态交互权重的特征选择方法 IWFS，可以动态的影响到候选特征与类别变量的交互关系，同时，以对称不确定性的互信息来衡量相关关系，其评价标准如下：

$$J_{\mathrm{IWFS}}\left(X_i\right)=\left(IW\left(X_i,X_j\right)+1\right)\times SU\left(X_i,Y\right)$$

其中，$IW\left(X_i,X_j\right)=1+\dfrac{I\left(X_i;X_j;Y\right)}{H\left(X_i\right)+H\left(X_j\right)}$ 定义为动态交互因子，并且得到了 $0\le IW\left(X_i;X_j\right)\le 2$ 的结论。进一步分析后可以判断，若 $0\le IW\left(X_i;X_j\right)\le 1$，则两个特征之间为冗余的关系；若 $1\le IW\left(X_i;X_j\right)\le 2$，则两个特征之间为冗余的关系。$SU\left(X_i,Y\right)$ 代表的是对称不确定度，其定义为 $SU\left(X_i,Y\right)=2\times\dfrac{I\left(X_i;Y\right)}{H\left(X_i\right)+H\left(Y\right)}$，同时这也是一种标准化互信息的定义，在部分文献中，也被用 $NMI\left(X_i,Y\right)$ 来表示。

胡等人[13]提出了动态相关和联合互信息最大化算法(DRJMIM)，其评价标准如下：

$$J_{\mathrm{DRJMIM}}\left(X_i\right)=\min_{X_j\in S}\left(I\left(X_i;X_j;Y\right)\right)\times I\left(X_i;Y\right)+2\times\frac{I\left(X_i;Y\mid X_j\right)-I\left(X_i;Y\right)}{H\left(X_i\right)+H\left(Y\right)}\times I\left(X_j;Y\right)$$

高等人[14]将已选特征动态变化项与“最大相关最小冗余”原则相结合，采用先全局函数上首先选择最小的冗余特征，再选择最相关的特征的方式来进行特征选择，其评价标准是：

$$J_{\mathrm{NDCSF}}\left(X_i\right)=\arg\max_{X_i\in F-S}\left(\min_{X_j\in S}\left(I\left(X_i;Y\mid X_j\right)+I\left(X_j;Y\mid X_i\right)-I\left(X_i;X_j\right)\right)\right)$$

4. 本文算法

从对上述基于信息论的特征选择算法的分析来看，大多数的特征选择算法仅根据相关性和类无关特征的冗余性，过于依赖类的特征。这些算法存在两个明显的缺点，一个是计算相关性时，忽略了类别与特征之间因数量大而存在偏差，二是对特征冗余的高估，无法处理好特征间的高阶关系。

针对于上面总结出的两大问题，分别提出解决办法，综合考虑后，提出一种新的特征选择方法：

第一，针对互信息在多值目标与特征之间存在的偏差问题，使用对称不确定性度量的互信息来解决。对称不确定性度量的互信息也是一种标准化的互信息，其表示如下：

$$NMI\left(X_i;Y\right)=2\times\frac{I\left(X_i;Y\right)}{H\left(X_i\right)+H\left(Y\right)}$$

当特征间具有复杂的交互关系时，单纯的计算互信息来衡量特征与目标变量间的相关关系，忽略其交互的信息在总的不确定度下的比重，也可能会高估某些特征的重要性，而通过考虑对称不确定度的互

信息，是一种很好的解决方案。

第二，针对特征冗余的高估，无法处理好特征间的高阶关系的问题，所有考虑到冗余项的基于信息论的特征选择方法都需要计算包含已选特征子集的互信息或条件互信息，如 $I(S,Y)=I(X_1,X_2,\cdots,X_k;Y)$。其中的计算涉及到高维的概率密度函数，由于样本的限制，所有关于已选特征子集的互信息都不能得到精确计算，只能采用逼近的方法近似计算。经典的特征选择算法都是在使用多个低维互信息公式的相加来近似表示高维互信息的计算，近似过程中不可避免地会产生重复计算而过高的估计某些特征所含的信息，造成冗余。

所以要避免使用线性累和的方法计算冗余度，可以通过极端值来代替平均累积和。可以用 $I(X_i,S,Y)$ 来表示待选变量 X_i 与已选特征集合 S 和目标变量 Y 之间的交互信息，$I(Y;X_i\mid S)$ 来表示已知已选特征集合 S 时，待选变量 X_i 和目标变量 Y 之间的条件互信息。通过信息论的知识可得，

$$I(Y;X_i\mid S)\le\min_{X_j\in S}\left(I(Y;X_i\mid X_j)\le\frac{1}{|S|}\sum_{X_j\in S}I(Y;X_i\mid X_j)\right)$$

$$I(X_i,S,Y)\le\min\left\{I(X_i;S),I(S;Y),I(X_i;Y)\right\}$$

极端准则方法比使用平均原理的方法产生的冗余更少，并且在某些特殊情况下可以获得更好的结果。

我们可以通过交互信息与条件互信息来度量待选特征、已选特征子集和类别变量之间的冗余。首先，$I(X_i;X_j)$ 度量的是待选特征与一个已选特征之间的冗余，$I(X_i;X_j\mid Y)$ 度量的是已知类别变量 Y 时，待选特征 X_i 与已选特征 X_j 之间的冗余，$I(X_i;X_j\mid Y)$ 与 $I(X_i;X_j)$ 的值越大，说明候选特征与已选特征的关系越紧密，候选特征与已选特征越冗余。采用极端准则方法来降低对冗余的过高估计。$I(X_i;Y\mid X_j)$ 与冗余的差越大，代表对分类的结果帮助越大。

所以基于最大相关最小冗余原则的特征选择方法 NMI-JMI 的筛选标准为：

$$J_{\mathrm{NMIJMI}}(X_i)=NI(X_i,Y)-\max\left\{I(X_i;Y\mid X_j)-I(X_i;X_j\mid Y)\right\}$$

因此，NMIJMI 算法可以看作是一个过滤式特征选择算法。我们将它命名为 NMIJMI 是因为其结合了标准化互信息 NMI 与联合互信息 JMI 的原理。

算法的伪代码

输入：原始数据集 M，其中包括所有特征 $F=X_1,X_2,X_3,\cdots,X_n$ 和类别变量 Y；指定的特征子集阈值 K

输出：选择出的特征子集 S

1: 初始化：$S=\varnothing$，$k=1$

2: for $i=1$ do

3: 计算每个候选特征与类别变量之间的对称不确定度 $NMI(X_i;Y)$

4: end for

5: 从 $NMI(X_i;Y)$ 中选择最大的值，将令 $NMI(X_i;Y)$ 最大时的 X_i 记作 $X_{selected}$

6: $S=S\cup X_j$

7: $F=F-X_{selected}$

8: while $k<K$ do

9: for $X_i\in F$ do

10: 计算 $J_{\mathrm{NMIJMI}}(X_i)$

11: end for

12: 从 $J_{\text{NMIJMI}}(X_i)$ 中选择最大的值，将令 $J_{\text{NMIJMI}}(X_i)$ 最大时的 X_i 记作 $X_{selected}$

13: $S = S \cup X_j$

14: $F = F - X_{selected}$

15: $k = k + 1$

16: end while

5. 实验结果与分

为了说明所提出方法的性能，我们在十一个基准数据集上将其与六种基准方法 mRMR、CMIM、JMI、NMIFS、MRI 和 JMIM 进行了比较。由于本文的对比实验采用的是公开的标准数据集，所以不需要验证特征与类别变量的相关性。

这些数据集来自 UCI 机器学习库和 scikit 特征选择库，实例的大小从 77 到 7797 不等，特征数量从 18 到 5469 不等，类的数量从 2 到 26 不等。这些数据集来自不同的应用领域，其中 WDBC 和 DLBCL 是生物学数据，Movement 为视频处理数据，COIL20、ORL、faces 为人脸图像数据，Isolet 为音频处理数据，semeion 为手写图像数据，vehicle 是车辆数据，CNAE9 是文本数据，sonar 是声呐数据集，这些数据集的特点是如表 1 所示：

Table 1. Eleven public data sets
表 1. 十一个公共数据集

NO.	数据集	特征数	样本数	类别数	来源
1	vehicle	18	846	4	UCI
2	WDBC	30	569	2	UCI
3	sonar	60	208	2	UCI
4	movement-libras	90	360	15	UCI
5	semeion	256	1593	10	UCI
6	isolet	617	7797	26	UCI
7	CNAE9	856	1080	9	UCI
8	COIL20	1024	1440	20	UCI
9	faces	2914	1348	8	scikit
10	ORL	4096	400	40	scikit
11	DLBCL	5469	77	2	UCI

以上数据集都被广泛地用于各种特征选择算法中，所以本文使用这十一个数据集作为研究对象。虽然这十一个数据集都已经被从最开始的原始数据转为了数值型数据集，但是在进行特征选择之前，进行离散化处理。通过观察数据集的具体数据，本文选取等距离散法作为对连续型数据的离散化方法。为了便于对互信息的计算，决定将所有的数据划分到 20 个不同的区间并用区间的中点值对区间内的数据重新赋值。在本次实验中，我们对数据使用 10-折交叉验证划分训练集和验证集，实现对分类器的训练及验证过程，使用的是经过筛选得到的特征子集中的原始数据。为了消除量纲对结果的负面影响，使得所有特征在数值上具有可比性，对每列特征采用 min-max 标准化方法进行标准化处理。

如果数据集的特征数的一半超过 50，选择，50 个特征，如果数据集的特征数小于 100，则选择该数据集的所有特征数的一半。将不同特征选择算法下得到的特征子集在同一个分类器上学习，将获得的分类性能进行比较。表 2~4 为 7 个算法分别在 KNN 分类器(其中 K 取 3)、SVM 分类器、RF 分类器上的宏平均 F1 值。表中的宏平均 F1 值为各个算法 10 折交叉验证在各个分类器上获得的宏平均 F1 值的平均值。其中符号“±”后面的值为宏平均 F1 值的标准偏差。并且我们通过利用 t 检验方法对实验结果进行显著性测试，设置统计 P 值小于 0.05。符号“a”表示使用 t 检验之后，NMIJMI 在该数据集显著优于当前比较算法；符号“b”表示 NMIJMI 在该数据集显著劣于当前比较算法。最后一行“W/T/L”统计 NMIJMI 显著优于/不显著/显著劣于比较算法的数据集个数。同时使用粗体标示出每个数据集在该分类器下最高的宏平均 F1 值的平均值。

Table 2. F1_macro of seven algorithms on 3NN
表 2. 七种算法在 3NN 分类器上的宏平均 F1 值

数据集	NMIJMI	mRMR	CMIM	JMI	NMIFS	MRI	JMIM
vehicle	72.11 ± 3.02	68.91 ± 2.78a	68.21 ± 3.67a	61.82 ± 4.71a	61.46 ± 4.63a	68.55 ± 3.99a	63.72 ± 3.69a
WDBC	96.83 ± 2.07	93.85 ± 2.86a	95.26 ± 3.14a	96.3 ± 1.85	87.33 ± 3.56a	95.43 ± 2.26a	93.85 ± 3.86a
sonar	72.12 ± 9.93	56.67 ± 17.19a	69.24 ± 11.86a	70.26 ± 15.17a	53.29 ± 12.62a	59.57 ± 12.20a	72.07 ± 13.53
movement-libras	77.22 ± 12.90	70.28 ± 13.61a	74.45 ± 3.33a	76.71 ± 2.45	86.56 ± 2.13b	71.11 ± 16.44a	70.56 ± 14.81a
semeion	87.2 ± 3.44	73.63 ± 3.25a	72.75 ± 6.37a	84.5 ± 5.45a	83.5 ± 5.72a	77.33 ± 2.62a	60.45 ± 2.78a
isolet	89.58 ± 2.18	71.99 ± 2.40a	72.22 ± 13.15a	76.94 ± 12.55a	70.56 ± 13.56a	75.8 ± 2.51a	53.16 ± 2.55a
CNAE9	84.35 ± 3.60	81.2 ± 2.69a	81.3 ± 4.10a	82.59 ± 3.53a	83.33 ± 3.77	83.33 ± 3.07a	66.94 ± 3.29a
COIL20	94.17 ± 3.19	83.47 ± 6.34a	82.57 ± 8.21a	90.42 ± 6.78a	86.32 ± 6.00a	86.32 ± 5.12a	82.15 ± 6.65a
faces	62.32 ± 3.01	27.97 ± 2.39a	46.37 ± 2.95a	61.94 ± 3.92	29.15 ± 3.37a	22.86 ± 3.45a	38.95 ± 2.43a
ORL	90.25 ± 3.60	73.5 ± 8.08a	72.75 ± 6.37a	84.5 ± 5.45a	83.5 ± 5.72a	63.5 ± 6.34a	58.25 ± 7.25a
DLBCL	100 ± 0.00	79.11 ± 9.21a	96.07 ± 8.21a	88.57 ± 15.28a	73.75 ± 15.59a	80.54 ± 9.02a	88.39 ± 12.25a
W/T/L		11/0/0	11/0/0	8/3/0	9/1/1	11/0/0	10/1/0

表 2 是六种对比算法和本文提出的 NMIJMI 算法在 KNN 分类器上的宏平均 F1 值的平均值的比较，由表中结果可以看到，NMIJMI 方法在除了 movement-libras 数据集外的 10 个数据集上都取得了最高的宏平均 F1 值的平均值，在 DLBCL 数据集上取得了 100%的分类准确率，在 7 个数据集上的分类准确率超过了 80%。并且与 mRMR、CMIM、MRI 相比，在所有的施压数据集上都取得了显著优势，与 mRMR、CMMIM、JMI、NMIFS、MRI 和 JMIM 这六个对比算法相比，NMIJMI 取得显著优势的数据集个数分别为 11、11、8、9、11、10，同时仅在 movement-libras 数据集上显著弱于 NMIFS 方法。总的来说，利用 NMIJMI 算法筛选得到的特征子集能够在 3NN 分类器上使分类的精度得到有效地提高。也就意味着 NMIJMI 算法可以更好地保留原始数据中的有效信息。

表 3 是六种对比算法和本文提出的 NMIJMI 算法在 SVM 分类器上的宏平均 F1 值的平均值的比较，可知 NMIJMI 方法在 7 个数据集上取得了最高的宏平均 F1 值的平均值，在 6 个数据集上的分类准确率超过了 80%，分类的结果要比在 KNN 分类器上稍弱。同一数据集在这两种不同的分类器上得到的结果也互有优劣，如 WDBC 数据集、movement-libras 数据集和 ORL 数据集在 SVM 分类器上可以明显的得到比 KNN 分类器上更好的结果。与 mRMR、CMIM、JMI、NMIFS、MRI 和 JMIM 这六个对比算法相比，NMIJMI

取得显著优势的数据集个数分别为 10、8、7、8、8、11，比 KNN 分类器稍差。此次在 movement-libras 数据集上与于 NMIFS 方法结果相近。在 semeion 数据集得到的结果最差，显著弱于 6 个对比算法中的三个，低于最高的 NMIFS 算法 12%。在 WDBC 数据集上与 JMI 算法同时取得了最高的宏平均 F1 值的平均值，均为 97.33%，但是标准差要比 JMI 的小。总体上看，NMIJMI 算法在 SVM 分类器上表现良好，能够有效的提高分类性能。

Table 3. F1_macro of seven algorithms on SVM
表 3. 七种算法在 SVM 分类器上的宏平均 F 值

数据集	NMIJMI	mRMR	CMIM	JMI	NMIFS	MRI	JMIM
vehicle	69.92 ± 4.01	75.82 ± 4.30b	70.38 ± 2.10	70.63 ± 3.83	69.55 ± 4.57	76.69 ± 4.12b	63.11 ± 6.30a
WDBC	97.33 ± 1.96	96.2 ± 2.69a	96 ± 2.87a	97.33 ± 2.28	95.21 ± 2.38a	96.57 ± 2.54	94.29 ± 2.39a
sonar	71.61 ± 13.04	67.16 ± 16.27a	61.49 ± 14.99a	66.91 ± 16.80a	63.17 ± 14.24a	66.5 ± 16.99a	69.82 ± 17.36a
movement-libras	80.46 ± 12.87	77.86 ± 13.06a	80.94 ± 12.76	80.8 ± 11.37	80.01 ± 14.24	79.71 ± 13.79a	75.16 ± 14.30a
semeion	79.04 ± 3.13	77.65 ± 3.59a	78.6 ± 3.66	81.18 ± 2.93b	91.9 ± 1.92b	81.88 ± 2.65b	68.69 ± 3.83a
isolet	93.49 ± 1.66	78.62 ± 1.88a	74.48 ± 1.76a	87.33 ± 1.57a	81.46 ± 2.49a	84.48 ± 2.04a	58.82 ± 2.21a
CNAE9	75.6 ± 2.62	73.4 ± 3.89a	73.9 ± 3.82a	74.33 ± 3.52a	78 ± 3.46b	74.35 ± 3.68a	62.46 ± 5.67a
COIL20	95.93 ± 5.26	82.78 ± 6.28a	78.17 ± 5.97a	92.91 ± 6.20a	85.3 ± 6.49a	90.66 ± 6.60a	72.05 ± 4.83a
faces	68.29 ± 4.51	20.91 ± 2.16a	43.93 ± 4.95a	56.89 ± 5.16a	19.27 ± 2.67a	22.89 ± 3.28a	33.51 ± 5.53a
ORL	95.08 ± 5.87	79.71 ± 6.81a	78.93 ± 4.85a	89.71 ± 6.71a	89.96 ± 7.32a	70 ± 7.41a	65.76 ± 8.39a
DLBCL	94.97 ± 7.87	75.1 ± 22.39a	93.16 ± 10.56a	84.01 ± 24.12a	42.95 ± 1.17a	79.74 ± 17.55a	89.67 ± 12.13a
W/T/L		10/0/1	8/3/0	7/3/1	8/1/2	8/1/2	11/0/0

Table 4. F1_macro of seven algorithms on RF
表 4. 七种算法在 RF 分类器上的宏平均 F1 值

数据集	NMIJMI	mRMR	CMIM	JMI	NMIFS	MRI	JMIM
vehicle	73.06 ± 4.31	70.73 ± 4.37a	72.13 ± 4.21	70.96 ± 4.03a	68.21 ± 5.68a	70.88 ± 6.97a	68.14 ± 4.07a
WDBC	96.39 ± 3.71	95.45 ± 2.93	95.85 ± 3.27	96.05 ± 3.34	94.66 ± 4.46a	95.83 ± 3.02	93.77 ± 3.52a
sonar	74.14 ± 18.42	67.74 ± 17.17a	72.71 ± 19.80a	69.03 ± 20.24a	63.34 ± 20.27a	66.4 ± 18.18a	69.87 ± 20.83a
movement-libras	77.02 ± 10.04	76.18 ± 12.83	74.88 ± 12.49a	76.8 ± 11.78	77.82 ± 8.93	77.81 ± 10.41	71.51 ± 11.48a
semeion	81.44 ± 3.79	77.48 ± 2.82a	79.85 ± 3.37a	83.01 ± 3.04b	91.62 ± 2.45b	82.7 ± 3.55b	72.23 ± 0.96a
isolet	92.35 ± 2.34	77.17 ± 3.21a	75.94 ± 3.92a	86.46 ± 1.93a	80.25 ± 3.59a	82.88 ± 2.37a	66.37 ± 4.22a
CNAE9	87.82 ± 3.07	84.77 ± 3.07a	85.27 ± 2.88a	84.72 ± 4.34a	87.86 ± 3.17	87.25 ± 3.81	69.22 ± 3.88a
COIL20	98.22 ± 8.18	87.67 ± 10.05a	88.79 ± 6.52a	94.33 ± 6.10a	93.79 ± 5.25a	92.22 ± 7.77a	85.95 ± 4.49a
faces	55.89 ± 3.65	19.93 ± 3.21a	43.62 ± 3.25a	61.94 ± 3.92b	20.61 ± 3.15a	23.05 ± 3.26a	33.69 ± 2.22a
ORL	94.04 ± 5.68	78.54 ± 6.44a	80.38 ± 6.90a	90.67 ± 7.44a	90.63 ± 7.79a	63.5 ± 9.34a	77.14 ± 5.95a
DLBCL	91.53 ± 9.54	42.06 ± 14.69a	87.62 ± 12.14a	88.9 ± 20.62a	60.82 ± 11.59a	71.43 ± 20.04a	88.95 ± 13.35a
W/T/L		9/2/0	9/2/0	7/2/2	8/2/1	7/3/1	11/0/0

表 4 是七种算法在随机森林分类器上的宏平均 F1 值比较，可以看出，表现在随机森林分类器上表现最好的两个特征选择算法分别是 NMIJMI 算法和 NMIFS 算法。其中 NMIJMI 算法表现最好，在 8 个数据集上取得了最好的分类结果，在 6 个数据集上的分类精度超过了 80%，NMIFS 在余下的两个数据集中取得了最好的分类结果，且在 6 个数据集上分类精度超过 80%。JMI 与 MRI 结果相仿，JMI 算法略优于 MRI 算法。在 WDBC 和 movement-libras 数据集上，7 中算法都得到了相对较好的分类结果，在

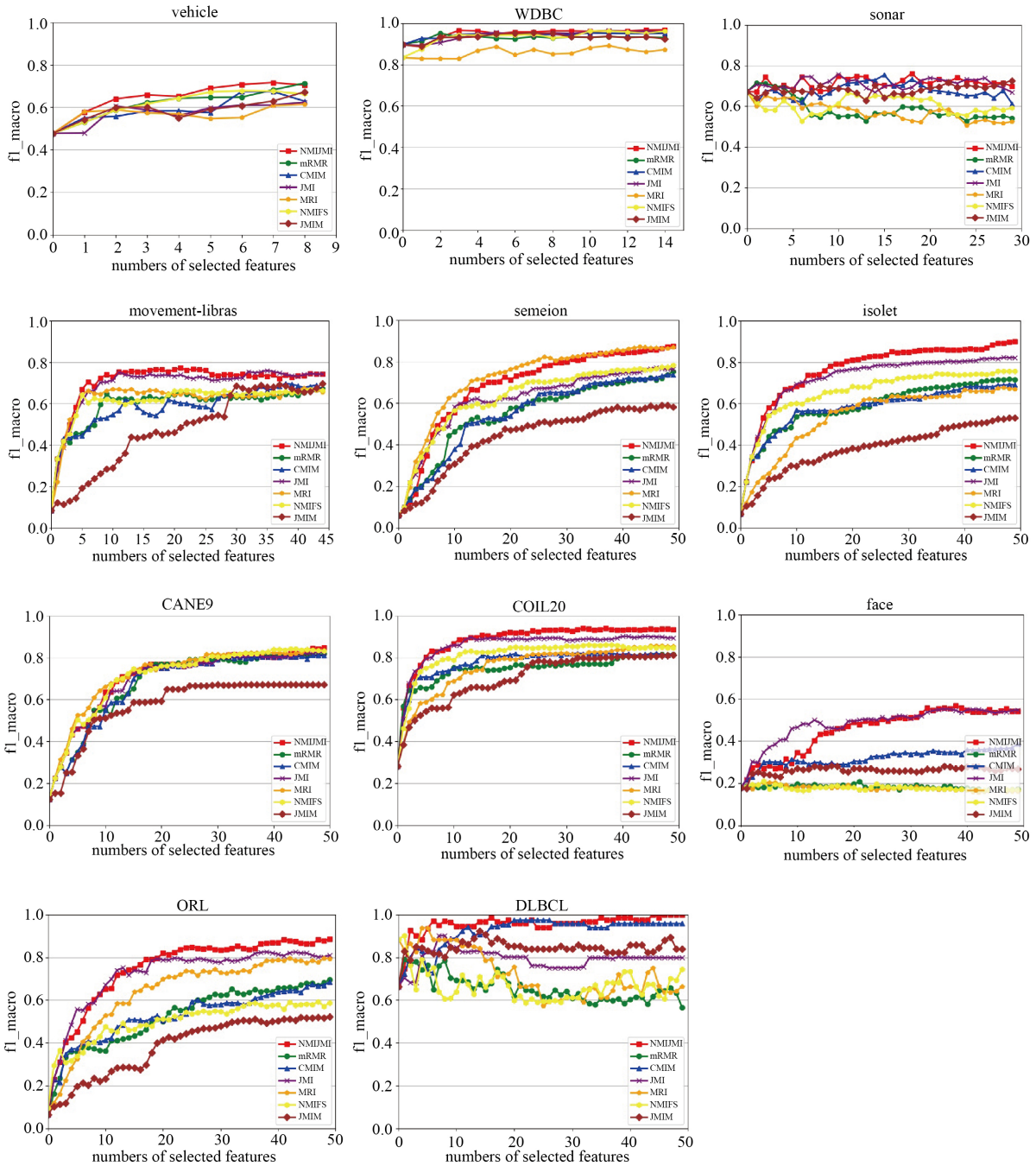

Figure 1. Line chart of F1_macro versus number of feature for 7 algorithms
图 1. 7 种算法的宏平均 F1 值随特征数变化折线图

movement-libras 数据集上，最好的结果与最差的结果间差距仅为 6.3%，其中五种算法结果的差距在 2%以内，在 WDBC 数据集上，最好的结果与最差的结果间差距不到 3%。也可以得出，NMIJMI 算法在随机森林分类器上也得到了良好的结果，可以认为其筛选效果优于其他 6 中特征选择算法。

为了能够清晰地说明不同的特征选择算法在选取不同数量的特征时的分类表现，本文给出这七中算法使用 KNN 分类器在这 11 个数据集下的分类宏平均 F1 值折线图，使结果更加具有说服力。如图 1 所示，图中的横坐标为特征的数量，纵坐标为宏平均 F1 值。

6. 结论

高维数据代表着数据中包含大量的信息，其中很多信息属于冗余信息和噪声信息，这些信息的存在其必然会导致数据的复杂性，加大数据分析与处理的难度。高维数据中，有效信息只是少数，如何保留有效信息，剔除无效的、冗余的信息以及无关的噪声信息，是统计学领域的热点问题，对机器学习、模式识别等领域有着重要意义。降维是处理高维数据问题的重要手段，而特征选择是一项重要的降维方式。本文针对高维复杂数据的特征选择展开了讨论与研究，重点研究内容是基于信息论的过滤式特征选择方法，目的是尽可能地保留相关特征，去除冗余的和无关的特征，达到降维的效果，从而提高学习算法的性能，同时降低计算成本，能够当作数据预处理的一个步骤，对数据分析与挖掘起到积极作用。文章依据最大相关最小冗余原则，利用互信息和条件互信息来综合度量特征的相关性和冗余性，着重对特征冗余性进行改进，并结合已选特征动态变化项，提出了一种非线性的过滤式特征选择算法。实验结果表明新的特征选择算法能够在某些数据集上取得良好的分类效果，选择出能提高分类器性能的特征子集。

参考文献

[1] Guyon, I. and Elisseff, A. (2003) An Introduction to Variable and Feature Selection. *JMLR*, **3**, 1157-1182.

[2] Baranauskas, J.A. and Netto, S.R. (2017) A Tree-Based Algorithm for Attribute Selection. *Applied Intelligence*, **48**, 821-833. https://doi.org/10.1007/s10489-017-1008-y

[3] Kohavi, R. and John, G.H. (1997) Wrappers for Feature Subset Selection. *Artificial Intelligence*, **97**, 273-324. https://doi.org/10.1016/S0004-3702(97)00043-X

[4] Chen, G. and Chen, J. (2015) A Novel Wrapper Method for Feature Selection and Its Applications. *Neurocomputing*, **159**, 219-226. https://doi.org/10.1016/j.neucom.2015.01.070

[5] Zhou, H.F., Guo, J. and Wang, Y.H. (2016) A Feature Selection Approach Based on Interclass and Intraclass Relative Contributions of Terms. *Computational Intelligence and Neuroscience*, **2016**, Article ID: 1715780. https://doi.org/10.1155/2016/1715780

[6] Zhou, H.F., Guo, J. and Wang, Y.H. (2016) A Feature Selection Approach Based on Term Distributions. *SpringerPlus*, **5**, Article No. 249. https://doi.org/10.1186/s40064-016-1866-5

[7] 赵晓群. 信息论基础及应用[M]. 北京: 机械工业出版社, 2015: 27-53.

[8] Brown, G., Pocock, A., Zhao, M.-J. and Lujan, M. (2012) Conditional Likelihood Maximisation: A Unifying Framework for Information Theoretic Feature Selection. *Journal of Machine Learning Research*, **13**, 27-66.

[9] Wang, J., Wei, J.M., Yang, Z., *et al.* (2017) Feature Selection by Maximizing Independent Classification Information. *IEEE Transactions on Knowledge and Data Engineering*, **29**, 828-841. https://doi.org/10.1109/TKDE.2017.2650906

[10] Fleuret, F. (2004) Fast Binary Feature Selection with Conditional Mutual Information. *Journal of Machine Learning Research*, **5**, 1531-1555.

[11] Bennasar, M., Hicks, Y. and Setchi, R. (2015) Feature Selection Using Joint Mutual Information Maximisation. *Expert Systems with Applications*, **42**, 8520-8532. https://doi.org/10.1016/j.eswa.2015.07.007

[12] Zeng, Z.L., Zhang, H.J., Zhang, R. and Yin, C.X. (2015) A Novel Feature Selection Method Considering Feature Interaction. *Pattern Recognition*, **48**, 2656-2666. https://doi.org/10.1016/j.patcog.2015.02.025

[13] Hu, L., Gao, W., Zhao, K., *et al.* (2017) Feature Selection Considering Two Types of Feature Relevancy and Feature Interdependency. *Expert Systems with Applications*, **93**, 423-434. https://doi.org/10.1016/j.eswa.2017.10.016

[14] 高万夫, 张平, 胡亮. 基于已选特征动态变化的非线性特征选择方法[J]. 吉林大学学报: 工学版, 2019(4): 8.