文章编号：1672-8785(2021)05-0033-06

基于零亏损冗余删减策略的最优光谱特征选择算法

吕子敬　张　鹏　刘志明　张志辉　韩　强
（中电科仪器仪表有限公司，山东 青岛　266555）

摘　要：为了以最小的代价筛选出最优的光谱特征，从信息亏损角度提出了一种 Filter 型光谱特征选择算法。该算法依据联合互信息的大小对特征进行排序，并采用一种零信息亏损原则对排序后特征全集中的冗余特征进行判断和删减。这样既能获得一个规模较小并可表现整个原始光谱特征的最优光谱特征子集，又减小了由冗余特征删减带来的信息亏损。

关键词：特征选择；联合互信息；零信息亏损；冗余特征；特征子集

中图分类号：TH744　　**文献标志码**：A　　**DOI**：10.3969/j.issn.1672-8785.2021.05.006

An Optimal Spectral Feature Selection Algorithm Based on Zero Loss Redundancy Reduction Strategy

LV Zi-jing，ZHANG Peng，LIU Zhi-ming，ZHANG Zhi-hui，Han Qiang
（*China Electronics Technology Instruments Co.，Ltd.，Qingdao 266555，China*）

Abstract：In order to select optimal spectral features at the minimum cost，a Filter type algorithm for spectral feature selection is proposed from the perspective of information loss. The algorithm sorts the features according to the joint mutual information size and uses a zero information loss principle to judge and reduce the redundant features of the sorted feature set. In this way，a small and optimal subset of spectral features can be obtained，which can represent the whole original spectral features，and the information loss caused by the deletion of redundant features can be reduced.

Key words：feature selection；joint mutual information；zero information loss；redundancy feature；feature subset

0　引言

随着光学技术的发展，光谱信息的数据量不断增大，人们对光谱数据的分析也不断加深。作为一种挖掘光谱数据的关键手段，特征选择对光谱数据处理的精确度、特征提取的效率以及物质的定性分析都有重要影响[1]。特征选择研究始于 20 世纪 60 年代。它包含在整个学习算法当中，并发生在分类学习操作之前。按照特征子集的产生过程对

收稿日期：2020-11-09

作者简介：吕子敬(1985-)，男，山东青岛人，工程师，主要研究方向为光谱分析算法。
E-mail：570824026@qq.com

最终使用它的归纳学习算法的依赖性进行分类。特征选择包括 Embedded 型、Filter 型和 Wrapper 型三类[2]。对于 Filter 型算法来说，具体的特征选择过程相对独立。以特征评价为标准，它的选择方法分为两种：基于特征子集评价和基于单一评价。Yu L 等人提出的 FCBF 算法就是一种两阶段评价的特征选择算法[9]。该算法提出了一种新的互信息度量 SU，并将其用于对特征集进行排序和筛选。然后利用一个近似马尔科夫毯的定义来删减所得特征子集中的冗余特征[3]。该算法有较高的执行效率。但是近似马尔科夫毯的不严格性使它在一些情况下不能获得较好的结果。综上考虑，本文提出了一种新的 Filter 型特征选择算法——基于零亏损冗余删减策略的最优光谱特征选择算法(NLPS)。该算法可应用于红外光谱特征提取过程，并可更精确地获取最能表现原始光谱信息的最优特征数据信息，从而为后续的物质定性分析提供依据。

1 NLPS 算法

NLPS 算法首先按照联合互信息由低到高对特征进行排序。然后以零信息亏损为原则对光谱特征进行冗余判断。这样做既保留了互信息最大的特征(保证光谱特征子集尽量小)，又在零信息亏损的情况下删除了冗余特征。下面引入等信息特征子集评价准则，然后在此基础上提出 NLPS 算法，并证明其结果为等信息光谱特征子集，最后给出整个光谱特征选择算法的伪代码。

1.1 等信息特征子集评价准则

为了确保两个阶段筛选出的光谱特征子集具有较高的代表性，引入了等信息特征子集评价准则。首先对贝叶斯错误率进行量化分析。E 表示类标签，M 表示整个数据集，$\{T_1, T_2, \cdots, T_n\}$ 表示特征集 T 的特征合集。贝叶斯错误率上界[3]为

$$P_{bayes} \leqslant \frac{1}{2}(H(E) - JS[p(T_1, T_2, \cdots, T_n)|E]) \tag{1}$$

式中，JS 表示 JS 散度，H 表示信息熵。

$$\begin{aligned} &JS[p(T_1, T_2, \cdots, T_n)] \text{ 满足} \\ &JS[p(T_1, T_2, \cdots, T_n|E)] = \\ &H[\sum_{e\in E} p(e)p(t_1, t_2, \cdots, t_n|e)] \\ &-\sum_{e\in E} p(e)H(p(t_1, t_2, \cdots, t_n|e)) \\ &= H(T) - H(T|E) \\ &= I(E;T) \end{aligned} \tag{2}$$

因此，式(1)可表示为

$$P_{bayes} \leqslant \frac{1}{2}(H(E) - I(E;T)) \tag{3}$$

对于特征集 T 而言，$S, T \subset \boldsymbol{T}$。根据联合互信息的链式方法[4]，可得

$$I(E;T|S) = I(E;S,T) - I(E;S) \tag{4}$$

由于条件互信息具有非负特性，且 I(E;T|S) 满足该特性，因此 I(E;T|S) 的值不小于零，可得

$$I(E;S) \leqslant I(E;S,T) \leqslant \cdots \leqslant I(E;T) \tag{5}$$

由式(3)和式(5)可得

$$\begin{aligned} P_{bayes} &\leqslant \frac{1}{2}(H(E) - I(T;E)) \\ &\leqslant \frac{1}{2}(H(E) - I(E;S)) \end{aligned} \tag{6}$$

从式(1)中可以看出，当 I(E;S) 增加时，(H(E)−I(E;S))/2 将更加接近贝叶斯错误率的实际值，即特征子集 S 的分布接近特征全集 T 所表示样本集的分布。由此给出等信息特征子集的定义：设 G⊂F，若满足 I(C;G)=I(C;F)，则称 G 为 F 的一个等信息特征子集。

显然，为了确保光谱特征子集的准确性，利用基于最小联合互信息亏损的光谱特征选择策略筛选出的光谱特征子集应该是一个等信息特征子集。为了得到等信息特征子集，提出以下性质。

性质 1.1：设当前已选特征集为 U_k，T 为特征全集。若 $\forall T \in T \backslash U_k$，满足 $I(E;T|U_k)=0$，则 $I(E;U_k)=I(E;T)$。

证明过程如下：

结合式(4)，有

$$I(E;T) - I(E;U_k) = I(E;T \backslash U_k|U_k) \tag{7}$$

对于变量集 P、Q 和变量 N、E，条件互信息[5]有以下性质：

$$I(P,N;E|Q)\leqslant I(P;E|Q)+I(N;E|Q) \tag{8}$$

结合式(7)，有

$$I(E;T\backslash U_k|U_k)\leqslant I(E;T\backslash(U_k\cup\{T_1\})|U_k)+I(E;T_1|U_k) \tag{9}$$

$$I(E;T\backslash U_k|U_k)\leqslant I(E;T\backslash(U_k\{T_1,T_2\})|U_k)+I(E;T_1|U_k)+I(E;T_2|U_k) \tag{10}$$

$$\vdots$$

$$I(E;T\backslash U_k|U_k)\leqslant\sum_{i=1}^{m}I(E;T_i|U_k),\quad \mathrm{I(E;T\backslash U_k|U_k)}\leqslant 0 \tag{11}$$

式中，m 为集合 $T\backslash U_k$ 的特征数，$T_i\in T\backslash U_k$。由于条件互信息具有非负特性，可得

$$0\leqslant \mathrm{I(E;T\backslash U_k|U_k)}\leqslant 0 \Rightarrow \mathrm{I(E;T\backslash U_k|U_k)}=0 \tag{12}$$

结合式(1)可得

$$\mathrm{I(E;T)-I(E;U_k)}=0 \tag{13}$$

即

$$\mathrm{I(E;T)=I(E;U_k)} \tag{14}$$

利用性质 1.1 可以判断所选特征子集是否为等信息特征子集。

1.2　算法原理

为了使光谱特征子集的规模最小，需要保留联合互信息最大的光谱特征。基于冗余删减策略的最优光谱特征选择算法首先按照联合互信息从小到大对当前光谱特征集进行排序，然后对该特征集进行冗余判断和删减。这样可以将对联合互信息大的光谱特征进行的冗余判断放在最后，从而确保这些光谱特征不会在前期被删除。为了使光谱特征删减后的信息亏损尽可能小，需要对每个光谱特征删减后的互信息亏损进行分析，以决定是否删除该特征。也就是说，对于当前光谱特征子集 U_k 以及正在分析的特征 T_i，需要使 $I(E;U_k)-I(E;U_k\backslash T_i)$ 尽可能小。由性质 1.1 可知：

$$I(E;U_k)-I(E;U_k\backslash T_i)=I(E;T_i|U_k\backslash T_i) \tag{15}$$

于是将问题转化为要删除的光谱特征 T_i 必须满足 $I(E,T_i|U_k\backslash T_i)$尽可能小。由条件互信息的非负性得知 $I(E,T_i|U_k\backslash T_i)\geqslant 0$。因此当 T_i 满足 $I(E,T_i|U_k\backslash T_i)=0$ 时，$I(E;U_k)-I(E;U_k\backslash T_i)$取得最小值，互信息亏损为 0。此时 T_i 为显著冗余特征。然而在特征维数高、样本数量少的情况下，对于两个条件独立的特征而言，它们的条件互信息数值一般不会为 0，而是在一个区间内变化。为了降低这种偏差给实验数据造成的影响(不准确)，可以引入一个判定数据冗余的阈值 ε。如果 $I(E;T_i|U\backslash\{T_i\})<\varepsilon$，则判定 T 为一个非常显著的冗余特征[6-7]。下面证明采用该删减策略得到的光谱特征子集 U 为 S 的等信息特征子集。

设 U 为经过基于最大联合互信息的特征选择策略筛选出的特征集，r_i 与 r_{i+1}为先后删减的两个冗余特征，P 为删减 r_i 后的特征子集，Q 为删减 r_{i+1}后的特征子集。则 P、Q 满足关系 $P=Q\cup\{r_{i+1}\}$。在这里令 $P=\{r_1,r_2,\cdots,r_{i-1},r_{i+1}\}$，$Q=\{r_1,r_2,\cdots,r_{i-1}\}$，则存在关系 t_i 满足

$$r_i=t_i(r_1,r_2,\cdots,r_{i-1},r_{i+1}) \tag{16}$$

对于 r_{i+1}，存在关系 t_{i+1}满足

$$r_{i+1}=t_{i+1}(r_1,r_2,\cdots,r_{i-1}) \tag{17}$$

由式(16)和式(17)可得

$$r_i=t_i(r_1,r_2,\cdots,r_{i-1},t_{i+1}(r_1,r_2,\cdots,r_{i-1})) \tag{18}$$

由式(18)可知，r_i 对于特征子集 Q 也是冗余特征。即对于类标签 E，r_i 满足 $I(E,r_i|Q)=0$。

上面证明的问题可表述为“前阶段的冗余特征也是后阶段的冗余特征”。因此，设该删减策略获得的特征子集为 S，被删除的特征集 $R=\{r_1,r_2,\cdots,r_n\}$，类标签为 E，存在以下关系：

$$\forall r\in R,\ I(r;E|S)=0 \tag{19}$$

由性质 1.1 可以判定 $I(E;S)=I(E;U)$，即 S 是 U 的等信息特征子集。

1.3　算法的具体实现

1.3.1　算法伪代码

以下是 NLPS 算法的伪代码：

```
算法：NLPS
输入：原始光谱数据集 M(T,E)、阈值 ε
输出：表达原始光谱信息的特征子集 U
1. 初始化参数：U ←ϕ，a ← 1，b ← 1；
2. repeat
3.    Choose T∈T\U which minimizes I(E;T)；
4.      if I(E;T)＞ 0 then
5.         T_a←T；
6.         U←U⋃{T_a}；
7.         a←a+1；
8.      end
9. until T\U is NULL；
10. while b<a
11.      do
12.        if I(E;T_b\U\{T_b})=0 then
13.           U← U\{T_b}；
14.      end
15.      else
16.        if I(E;T_b\U\{T_b}) <ε then
17.           U← U\{T_b}；
18.        end
19.      end
20.      b←b+1；
21.   end
22. end
```

第 2～9 行基于联合互信息大小对光谱特征集进行升序排序，第 10～22 行为冗余特征的判断和删减。整个算法进行了 $O(|\boldsymbol{T}|)$次联合互信息的估计和 $O(|\boldsymbol{T}|)$次条件互信息的估计。

1.3.2 算法的复杂度分析

前面已经分析过，对每个光谱特征联合互信息估算的时间复杂度为 $O(2(N+r))$。因此对光谱特征全集联合互信息估算的时间复杂度为 $|\boldsymbol{T}|O(2(N+r))$。排序过程采用复杂度为 $O(|\boldsymbol{T}|\log|\boldsymbol{T}|)$的算法。所以排序部分的时间复杂度为$|\boldsymbol{T}|O(2(N+r))+O(|\boldsymbol{T}|\log|\boldsymbol{T}|)$。在冗余删减部分，首先以前一阶段得到的光谱特征子集中的特征为关键字对样本集进行基数排序，时间复杂度为$O(|\boldsymbol{T}|(N+r))$。根据前面的分析，当计算每个光谱特征的条件互信息时，都需要遍历每个光谱特征集。因此计算每个特征的条件互信息所需的时间复杂度就是$O(N)$，整个冗余判断过程的时间复杂度为$|\boldsymbol{T}|O(N)+O(|\boldsymbol{T}|(N+r))$。整个算法的时间复杂度为

$$|\boldsymbol{T}|O(2(N+r))+O(|\boldsymbol{T}|\log|\boldsymbol{T}|)+O(|\boldsymbol{T}|(N+r))+|\boldsymbol{T}|O(N)$$

即

$$|\boldsymbol{T}|O(N+r)+O(|\boldsymbol{T}|\log|\boldsymbol{T}|)+O(|\boldsymbol{T}|(N+r))$$

其中，整个特征全集的特征个数用 T 表示，即原始光谱数据集的特征个数。被去除掉的光谱数据的冗余特征个数用 r 表示，算法执行过程中的迭代次数用 N 表示。

2 实验结果分析

为了获得较好的实验效果，在实验过程中采用了比较常用的基准数据集。它们分别是 DNA_All、Kr-vs-kp 和 Lung_Cancer，都来自 UCI 机器学习库[8]。同时本次实验将 NLPS 算法与 FCBF、ID3 以及 ReliefF 三种经典特征选择算法进行了比较，另外还选用了三种经典分类器：C4.5 决策树分类器、K 近邻(k-Nearest Neighbor，KNN)分类器和朴素贝叶斯分类器(Naive Bayes Classifier，NBC)。表 1 列出了所用数据集的特性数据。

表 1 所用数据集的特性数据

数据集	特征数	样本个数	类标签
DNA_ALL	3186	180	3
Kr-vs-kp	37	3196	2
Lung_Cancer	12600	203	5

为了准确地检测四种算法的性能，通过对比体现出 NLPS 算法的优越性。从表 1 中可以看出，所用数据集的特征维数与样本个数范围非常广泛。

2.1 实验平台与算法性能比较

Weka 是一种常用的算法分类性能比较平台[9]，其内部集成了经典的特征选择算法和分类器。比如需要用到的 ReliefF、ID3 与 FCBF 参照算法以及 KNN ($k=1$)、C4.5 与 NBC 分类器都可以在该平台中直接调用。使用 Java 编

写的 NLPS 算法和表 1 中的数据集也可以导入到该平台，并可用于分类性能对比实验，从而使实验过程简洁明了。因此，本文实验采用了该平台。

分类准确性是体现特征选择算法优劣的标准。因此，在实验过程中分别获取四种算法在数据集每个准确特征上的分类准确率，并通过对比它们的分类准确率曲线来评价算法优劣。四种选择算法在三种数据集中分类准确率的对比分别如图 1、图 2 和图 3 所示。

从实验结果中可以看出，对于不同的数据集，每种算法的表现都不相同。没有一种算法在所有数据集上的表现都是最优的。这也比较符合 No Free Lunch 理论[9]。对于样本个数最少的 DNA_ALL 数据集而言，这四种算法在数据冗余特征处理过程中的分类准确率相差不大。随着样本数的不断增加，NLPS 算法的分类准确度逐渐地略微高于其他三种算法。但是分类准确度数值曲线不是完全平缓地增长，而是有一定的跳变。这说明分类准确度受到样本个数的影响，但不是很明显。Kr-vs-kp 数据集的样本个数最多。随着样本个数的不断增加，四种算法的分类准确度数值曲线也完全平缓增长，没有任何数值跳变。NLPS 算法仍然能表现出优秀的分类效果，说明对样本个数很大的数据集而言，分类效果受特征冗余性影响。

当所选特征个数不断增加时，分类准确度的优势会增大，选择的特征作为一个整体具有更强的表达特性。NLPS 算法的分类准确度优势在特征数更多的 Lung_Cancer 数据集上也得到了较为深刻的体现。在特征数不断增大的过程中，数据集特征之间的冗余度、特征与类标签的相关度会对分类准确度产生重大影响。因此，特征数与样本个数这两个变量都能影响算法的分类准确度。总之，在这四种算法中，NLPS 算法具有较强的检索数据集相关特征、删除冗余特征以及最大限度保留数据特性的能力。

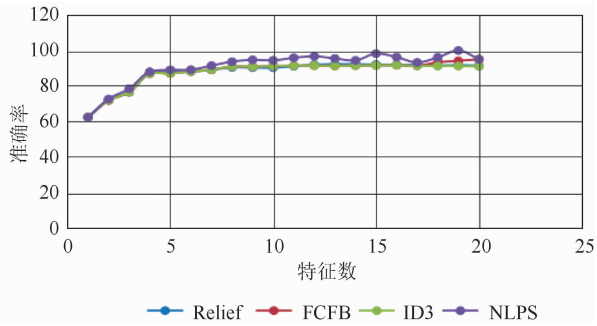

图 1 四种算法在 DNA_ALL 上的平均准确率曲线

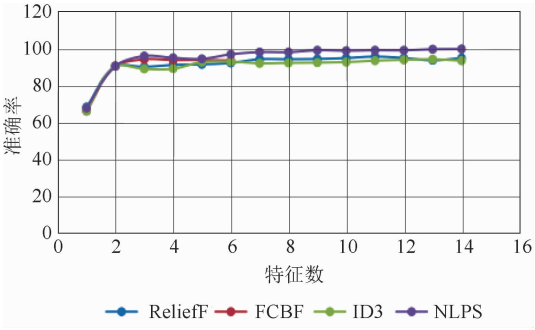

图 2 四种算法在 Kr-vs-kp 上的平均准确率曲线

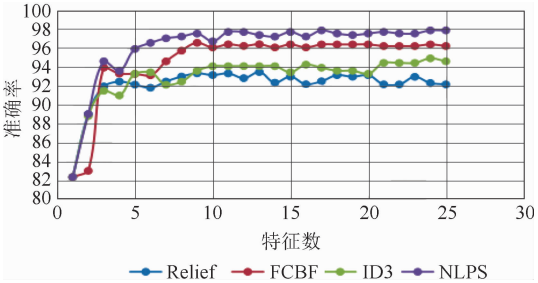

图 3 四种算法在 Lung_Cancer 上的平均准确率曲线

3 结束语

本文针对传统光谱特征选择算法中存在的一个普遍问题，提出了 NLPS 算法。它能够去除光谱数据集中的冗余特征，同时还可保留最能体现光谱特性的特征。针对特征与类的相关性、特征间的冗余性提出了新的信息论度量标准。在此基础上提出并实现了特征选择算法。利用三种基准数据集在 NBC、KNN 以及 C4.5 决策树三种经典分类器上进行了实验。实验结果证明了 NLPS 算法的优越性和有效性。尤其对于特征维数较高的数据集而言，NLPS 算法的相关特征搜索和冗余特征判别能力都明显优

于 FCBF、ID3 和 ReliefF 三种经典特征选择算法。

参考文献

[1] Quinlan J R. Induction of Decision Trees [J]. *Machine Learning*, 1986, **4**(2): 81–106.

[2] 崔自峰, 徐宝文, 张卫丰, 等. 一种近似 Markov Blanket 最优特征选择算法 [J]. **计算机学报**, 2007, **30**(12): 2074–2081.

[3] Miyahara K, Pazzani M J. Collaborative filtering with the simple Bayesian classifier [C]. Heidelberg: Pacific Rim International Conference on Artificial Intelligence, 2000.

[4] 张逸石, 陈传波. 基于最小联合互信息亏损的最优特征选择算法 [J]. **计算机科学**, 2011, **38**(12): 200–205.

[5] Hall M A. Correlation-based feature selection for discrete and numeric class machine learning [C]. Los Altos: 7th International Conference on Machine Learning, 2000.

[6] Ding C, Peng H. Minimum Redundancy Feature Selection from Microarray Gene Expression Data [C]. San Francisco: IEEE Computer Society Conference on Bioinformatics, 2003.

[7] 史衷植. **知识发现** [M]. 北京: 清华大学出版社, 2002.

[8] Li J. Divergence measures based on the Shannon entropy [J]. *IEEE Transactions on Information Theory*, 1991, **37**(1): 145–151.

[9] Yu L, Liu H. Efficient feature selection via analysis of relevance and redundancy [J]. *Journal of Machine Learning Research*, 2004, **5**(12): 1205–1224.