文章编号： 1672-8785(2016)01-0040-05

一种动态赋权红外光谱特征选择算法研究

吕子敬[2] 韩顺利[1,2] 张志辉[2] 刘 磊[2]

(1. 电子测试技术重点实验室，青岛 266555；
2. 中国电子科技集团公司第四十一研究所，青岛 266555)

摘 要：大规模的红外光谱数据集中存在大量无关冗余的特征。针对这一问题，提出了一种动态赋权红外光谱特征选择算法 (Dynamic Weight Infrared Spectrum Feature Selection Algorithm, MBDWFS)。该算法把对称不确定性度量标准与近似 Markov Blanket 相结合，以删除原始光谱数据集中无关冗余的特征，从而获取数据规模较小且最优的特征子集。通过与 FCBF 、 ID_3 和 ReliefF 三种经典特征选择算法的性能仿真对比试验，证明所提出的 MBDWFS 算法在整体分类性能上优于其他三种算法，用于红外光谱的物质分析领域时效果更好。

关键词：特征选择； Markov Blanket ； 动态赋权

中图分类号：TP181 **文献标志码**：A **DOI**: 10.3969/j.issn.1672-8785.2016.01.008

Study of a Dynamic Weighted Infrared Spectrum Feature Selection Algorithm

LV Zi-jing[1], HAN Shun-li[1,2], ZHANG Zhi-hui[1],LIU Lei[1]

*(1.Science and Technology on Electronic Test and Measurement Laboratory, Qingdao 266555, China;
2.The 41st Institute of CETC, Qingdao 266555, China)*

Abstract: There exist a large number of irrelevant and redundant features in large-scale infrared spectrum datasets. To solve this problem, a dynamic weighted infrared spectrum feature selection algorithm (MBDWFS) is proposed. The algorithm deletes the irrelevant and redundant features in an original spectrum dataset by combining the symmetric uncertainty metrics with Markov Blanket. Then, a smaller scale optimal feature subset is obtained. By comparison with three classical feature selection algorithms FCBF, ID_3 and ReliefF, it shows that the proposed MBDWFS algorithm is better than the above three algorithms in overall classification performance and is more suitable to be used in the field of material infrared spectrum analysis.

Key words: feature selection; Markov Blanket; dynamic weight

0 引 言

在当今的信息化时代，信息技术日新月异，计算机应用不断更新，红外光谱数据集的规模也随之不断扩大。由于光谱数据集中往往存在大量冗余信息，而这些冗余信息对机器学习算法的执行效率又会产生很大的影响，因此去除光谱数据集中的冗余信息就成为了特征选择算法所要解决的关键问题。特征选择算法的应用领域

收稿日期：2016–01–04
基金项目：国家重点实验室基金 (9140C12031150C12057)
作者简介：吕子敬 (1985-)，男，青岛人，硕士，助理工程师，主要从事光谱分析算法研究。
E-mail: lvzijing_1203@163.com

非常广泛，最突出的就是应用在物质分析领域，如特征选择算法可应用于红外光谱分析仪的红外光谱特征提取过程。该算法能更准确地选出原始光谱的目标信息，删除冗余信息，为后续的物质成分分析提供了强有力的保障。目前，特征选择算法有 Embedded 、 Filter 和 Wrapper 三种类型。Embedded 型特征选择算法被完全集成到特定学习算法的训练过程中， ID_3 决策树算法 [1] 是其典型代表。Wrapper 型特征选择算法用分类准确率作为度量准则， C4.5 算法 [2] 是其典型代表。与前两种算法不同， Filter 型算法的特征选择过程独立于具体的学习算法，特征选择是其分类的预处理操作。同 Embedded 和 Wrapper 算法比较起来， Filter 型特征选择算法具有计算代价小、效率高和适用范围广的特点。目前，国内外学者研究了很多高效 Filter 型特征选择算法，但普遍存在一个问题，一些作为特征集具有较强的表达特性但就其本身而言具有较弱表达特性的特征在选择过程中会被遗漏。其主要原因是，这些特征选择算法及其采用的信息度量方法忽略了特征的内部相关性。因此，本文提出了一种 MBDWFS 算法。

1 相关理论介绍

1.1 信息论度量

信息论中，信息熵和互信息是最基本的信息度量。它们能定量描述随机变量之间的不确定性。本文将介绍信息熵、条件信息熵、互信息和条件互信息。

假设有限离散型随机变量是 X, 变量 X 中所有可能的取值组合出现的概率分布密度函数为 $p(X)$ ，则 X 的信息熵 $H(X)$ 定义为 [3]

$$H(X)=-\sum_{i=1}^{n}p(x_i)\log p(x_i) \tag{1}$$

假设 X 、 Y 为两个有限离散型随机变量，变量 X 的条件信息熵用于描述当 Y 确定时 X 的不确定程度 [3] ：

$$H(X|Y)=-\sum_{y\in Y}p(y)\sum_{x\in X}p(x|y)\log p(x|y) \tag{2}$$

互信息用于描述两个变量间的相关性，给定两个有限离散型随机变量 X 、 Y, 互信息 $I(X;Y)$ 定义 [3] 为

$$I(X;Y)=\sum_{x\in X}\sum_{y\in Y}p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \tag{3}$$

由此可以看出，互信息也可以表示为 $I(X;Y)=H(X)-H(X|Y)$ 。

假设有限离散型随机变量是 Z, 则当 Z 确定时变量 X 和 Y 的条件互信息 [4] 可定义为

$$I(X;Y|Z)=H(X|Z)-H(X|Y,Z) \tag{4}$$

条件互信息可用于统计给定变量 Z 及当 Y 确定时 X 不确定性的减小程度。

1.2 新的信息度量方法

1.2.1 相关性度量

很多特征选择算法将互信息作为相关性的度量方法，这样存在一个显著问题，即此类算法倾向于选出互信息较大的特征，而没有考虑特征本身的不确定性。为了解决此问题，引入了对称不确定性度量 [4], 其定义为

$$SU(F,C)=2\frac{I(F;C)}{H(C)+H(F)} \tag{5}$$

1.2.2 冗余性度量

如果两个特征完全相关，这两个特征就是冗余的。因此，很多特征选择算法会把与已选特征集中高度相关的特征丢弃掉 。例如， mRMR 特征选择算法 [5] 就是利用最小冗余原则进行特征筛选，丢弃作为单一特征时与类较弱相关而作为特征集却具有较强表达特性的特征的。针对此问题，引入了新的度量方法，并将其用于度量特征 X 与特征 Y 的相对冗余度，其定义为

$$CSU(F,C)=2\frac{I(f_i;C|f_j)}{H(f_i)+H(C)} \tag{6}$$

将条件互信息引入对称不确定性度量，一方面考虑了特征间的冗余，另一方面又兼顾了特征与类标签的相关性。由于 $I(X,Y|C)\le Min\{H(X),H(Y)\}$, 因此 CSU 的值小于 1 。

1.3 近似 Markov Blanket

Markov Blanket 在 1996 年首次被 Koller 与 Sahami[5] 引入到特征选择中，该算法能有效删除无关冗余的特征。以下是 Markov Blanket 的一些基本概念：

定义 1 ：如果特征 $f_i \in N(N$ 为特征集) 且 $M_i \subset N(f_i \notin M_i)$,$M_i$ 就叫做 f_i 的 Markov Blanket，C 表示类标签，$P(N-M_i-\{f_i\},C|f_i,M_i)=P(N-M_i-\{f_i\},C|M_i)$ 。

虽然 Markov Blanket 可删除相关冗余的特征，但它对维数比较高的数据集进行空间搜索的时间复杂度很高，这会严重影响算法的运行效率，因此需要通过引入近似 Markov Blanket 来解决这一问题。

定义 2： R 表示一种度量准则，假定存在特征 f_i 和 f_j ，C 表示类标签。当 $R(f_i;C)>R(f_j;C)$ 时，f_i 就被称为 f_j 的一个近似 Markov Blanket，$R(f_j;C|f_i)>R(f_j;C)$ 。

2 MBDWFS 算法

基于前面的相关性、冗余性以及近似 Markov Blanket 分析，提出了一种新的特征选择算法 MBDWFS 。该算法主要分为三个阶段：首先，计算所有特征和类标签的 SU ，找出最大的 SU 值作为初始权重，并把 SU 值最大的特征加入到已选特征集中，删除候选特征集中的此特征；其次，以新加入到已选特征集的特征作为条件，计算所有特征的 CSU 值，找出初始权重与 CSU 相乘后的最大值，并把该值作为新的权重，同时把新的最大权重值对应的特征选入已选特征集中，并将该特征从候选特征集中删除；最后，用近似 Markov Blanket 删除已选特征集中冗余性较高的特征，直到候选特征集为空。表 1 是用伪代码描述的算法。

对于表 1 中的伪代码，可从整体上描述 MBDWFS 算法的时间复杂度：对由 n 个特征组成的候选特征集 S 来说，计算 SU 值和 CSU 值的时间开销都为 $O(n)$，所需的迭代次数为 n-d 次 (d 为候选特征集中删除的特征数目)，每次迭代需要计算 CSU 为 n-k 次，其中 k 为已选特征集的特征数目，此时，总体的时间复杂度为 (O(n2)-O(nd))；在最坏情况下即 d 为 0 时，该算法的总体时间复杂度为 $O(n^2)$，但在一般情况下，总体的时间复杂度会很低。

表 1 用伪代码描述的算法

输入：候选特征集 $S=\{f_1,f_2,\cdots,f_n,C\}$	
输出：最优特征子集 S_{best}	
1	begin
2	Initialize relative parameters: $S_{best}\leftarrow\phi$, $W\leftarrow\phi$, $F\leftarrow\{f_1,f_2,\cdots,f_n\}$;
3	foreach $f\in F$ do
4	$w(f)=SU(f,c)$
5	end
6	choose the feature $f_{max}\in F$ which maximizes $w(f)$
7	$W=W\cup\{f_{max}\}$
8	$F=F/\{f_{max}\}$
9	repeat
10	foreach $f\in F$ do
11	$w(f)=w(f)\times CSU(f,C\|f_{max})$
12	end
13	choose the feature $f\in F$ which maximizes $w(f)$
14	$W=W\cup\{f\}$
15	$F=F/\{f\}$;
16	if $(CSU(f;C\|f_i)>SU(f;C)$ do
17	delete f in W
18	end
19	end until F is NULL
20	$S_{best}=W$
21	end

3 实验验证

为了对 MBDWFS 算法与其他三种经典特征选择算法 (FCBF 、 ID_3 和 ReliefF) 进行性能比较，从 UCI 机器学习库 [5] 中选取了 DNA_All 、Kr-vs-kp 和 Lung_Cancer 三组数据集作为实验的基准数据集；另外，还选择了朴素贝叶斯分类器、 K 近邻分类器和 C4.5 决策树分类器 [5] 。实验数据集的详细描述见表 2 。

从表 2 中可看出，这些数据集无论是样本数还是特征维数，所涉及到的范围都非常广泛，能更全面地检测各算法的性能。

表 2　实验数据集描述

Datasets	Features	Instances	Classes
DNA_ALL	3186	180	3
Kr-vs-kp	37	3196	2
Lung_Cancer	12600	203	5

3.1　实验环境

该实验采用的平台是智能分析环境 Weka[6]。实验中所用到的 FCBF、ID_3 和 ReliefF 特征选择算法以及 NBC 分类器、C4.5 分类器、kNN 分类器均可以在 Weka 中直接调用，kNN 分类器中的 k 值为 1。表 2 中所有数据集的 MBDWFS 算法均采用 Java 语言编写，可在 Weka 环境中调用。

3.2　实验分析

分类准确率是评价特征选择算法优劣的最重要的指标。因此，我们在实验过程中采用十次十折交叉法来获取分类的准确率，并将分类准确率作为评价特征选择算法性能的指标。图 1、图 2 和图 3 分别表示 4 种选择算法在不同数据集中分类准确率的对比。

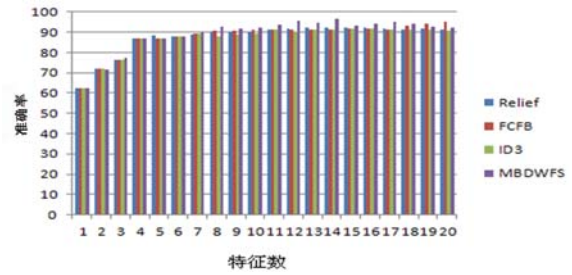

图 1　四种算法在 DNA_ALL 上的平均准确率曲线

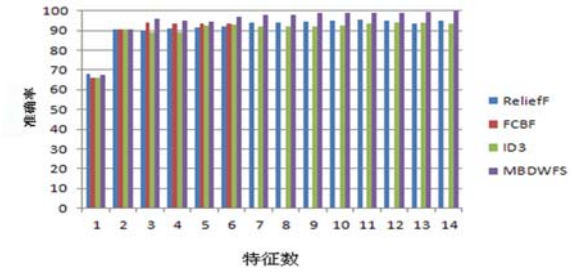

图 2　四种算法在 Kr-vs-kp 上的平均准确率曲线

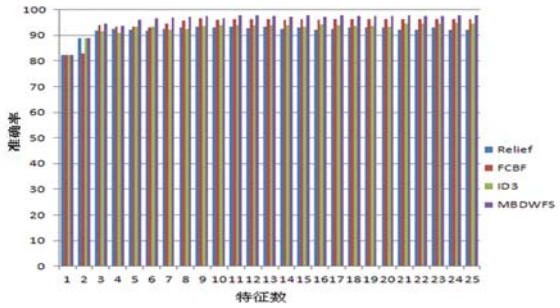

图 3　四种算法在 Lung_Cancer 日上的平均准确率曲线

从图 1、图 2 和图 3 的实验结果中可以看出，没有任何一种特征选择算法在所有的数据集上的分类准确率都是最优的，这正好也符合 No Free lunch 理论。对于 DNA_ALL 这种特征维数相对不算高的数据集来说，这四种算法优越性的差别不太明显，这在一定程度上说明了特征维数相对较低的数据集对于分类性能的影响不是很大。对 Kr-vs-kp 这种特征维数最高的数据集来说，与其他三种经典算法相比，MBDWFS 算法的分类效果更明显，性能更优越，这说明数据集的特征维数越高，数据集中的特征冗余性对分类的性能的影响就越大。随着所选择特征数目的不断增加，分类准确率增高优势的趋势越明显。从图 3 中看出，MBDWFS 算法的分类准确率优势对特征数较多的 Lung_Cancer 数据集而言较明显，这是因为随着特征数目的增多，特征间的冗余性以及作为整体对类标签的相关性便成为了影响分类准确率的重要因素，因而特征数目和特征维数的多少都会对分类的准确率产生至关重要的影响。总的来说，相比其他三种特征选择算法，MBDWFS 算法具有较强的相关特征搜索能力以及对冗余特征的判断能力。

4　结束语

针对红外光谱数据集中大量存在的冗余特征删除问题，提出了一种 MBDWFS 算法。介绍了特征选择算法的红外光谱物质分析领域的应用背景、算法的相关的理论基础、所提 MBDWFS 算法的思想和总体的时间复杂度，给出了 MBDWFS 算法、FCBF 算法、ID_3 算法和 ReliefF 算法的平均分类准确率这一辨别算法优劣性能的仿

真实验数据。从实验结果上看，所提出的 MBD-WFS 算法的相关特征搜索能力以及对冗余特征的判别能力都明显优于其他三种经典算法。

参考文献

[1] Zhang Y S,Zhang Z G,Liu K J,et al. An Improve Algorithm for Markov Blanket Discovery[J].*Journal of Computers*,2010, **5**(11):1755–1761.

[2] Hall M A. Correlation-based Feature Selection for Discrete and Numeric Class Machine Learning[C]//Proceedings of the 7th International Conference on Machine Learning, Los Altos: Morgan Kaufmann,2000,359–366.

[3] 毛勇，周晓波，夏铮，等. 特征选择算法研究综述 [J].**模式识别与人工智能**,2007,**20**(2):11–218.

[4] 王娟，慈林林，姚康泽. 特征选择方法综述 [J].**计算机工程与科学**,2005,**27**(12):68–71.

[5] Sotoca J, Pla F. Supervised Feature Selection by Clustering Using Conditional Mutual Information Based Distances[J].*Pattern Recognition*,2010,**43**(6):2068–2081.

[6] Yu L, Liu H. Efficient Feature Selection via Analysis of Relevance and Redundancy[J].*Journal of Machine Learning Research*,2004,**5**(12):1205–1224.

(上接第 39 页)

表 1 测试结果比较

样品	时域光谱峰值	频域光谱峰值频率 (THz)	折射率平均值	吸收系数峰值频率 (THz)
NO.1	1.91×10^{-3}	0.216 、 0.276 、 0.637	1.23	0.937
NO.2	1.84×10^{-3}	0.276 、 0.637	1.20	0.733
NO.3	1.11×10^{-3}	0.264 、 0.373 、 0.661 、 0.829	1.14	0.697 、 0.853 、 0.949
NO.4	3.43×10^{-3}	0.276 、 0.625	1.49	0.252 、 0.433 、 0.577 、 0.709

参考文献

[1] 张存林. **太赫兹感测与成像** [M]. 北京：国防工业出版社， 2008.

[2] 许景周，张希成. **太赫兹科学和技术** [M]. 北京：北京大学出版社， 2007.

[3] Fukunaga K, Hosako I, Picollo M, et al. Application Of Thz Sensing To Analysis Of Works Of Art For Conservation[C]. IEEE Topical Meeting on Microwave Photonics, 2010:147 – 150.

[4] Jackson J B, Mourou M R. Terahertz Time-domain Reflectometry Applied to the Investigation of Hidden Mural Paintings[C]. Lasers and Electro-Optics, 2008 and 2008 Conference on Quantum Electronics and Laser Science. CLEO/QELS 2008. Conference on. IET, 2008:1–2.

[5] Walker G C, Jackson J B, Giovannacci D, et al. Terahertz Analysis of Stratified Wall Plaster at Buildings of Cultural Importance across Europe[C]. Proc Spie, 2013, **8790**(5):417–437.

[6] Sfarra S, Ibarra-Castanedo C, Ambrosini D, et al. Non-Destructive Testing Techniques to Help the Restoration of Frescoes[J].*Arabian Journal for Science and Engineering*, 2014,**39**(5):3461–3480.

[7] Fukunaga K, Hosako I, Kohdzuma Y, et al. Terahertz Analysis of an East Asian Historical Mural Painting[J].*Journal of the European Optical Society - Rapid publications*, 2010, **5**(1):138–138.

[8] Candor é J C, Bodnar J L, Detalle V, et al. Nondestructive Testing of Works of Art by Stimulated Infrared Thermography[J].*European Physical Journal Applied Physics*, 2012,**57**(2):176–178.

[9] 李小霞, 邓琥, 廖和涛, 等. 室温下中药附子的太赫兹波谱分析 [J]. **激光与红外**, 2013, **43**(11):1282–1285.

[10] Dorney T D, Baraniuk R G, Dm. M. Material Parameter Estimation with Terahertz Time-domain Spectroscopy.[J]. *J Opt Soc Am A Opt Image Sci Vis*, 2001, **18**(7):1562–1571.