DOI：10.7652/xjtuxb201810019

利用近似马尔科夫毯的最大相关最小冗余特征选择算法

张俐，王枞，郭文明
（北京邮电大学可信分布式计算与服务教育部重点实验室，100876，北京）

摘要：针对高维数据集中冗余特征或无关特征降低机器学习模型分类准确率的问题，提出了一种基于近似马尔科夫毯的特征选择(nmRMR)算法。该算法首先利用最大相关最小冗余的准则进行特征相关性排序；采用近似马尔科夫毯算法对冗余特征或者无关特征进行删除，并最大程度地提高特征之间的相关性从而获得最优特征子集。在 UCI 的 8 个公开数据集上对比的实验结果表明：与 mRMR 算法相比，本文算法所选择出的特征子集数平均减少了 6.875 个，平均分类准确率提高了 0.78%；与 FullSet 算法相比，本文算法所选择出的特征子集数平均减少了 20.56 个，平均分类准确率提高了 1.88%；与 FCBF 算法相比，本文算法所选择出的特征子集数平均减少了 3.187 5 个，平均分类准确率提高了 0.825%；本文算法总体优于其他算法。

关键词：特征选择；特征相关；冗余特征；近似马尔科夫毯
中图分类号：TP181 **文献标志码**：A **文章编号**：0253-987X(2018)10-0141-05

A Feature Selection Algorithm for Maximum Relevance Minimum Redundancy Using Approximate Markov Blanket

ZHANG Li, WANG Cong, GUO Wenming
(Key Laboratory of Trustworthy Distributed Computing and Service Ministry of Education, Beijing University of Posts and Telecommunications, Beijing 100876, China)

Abstract: To solve the problem that redundancy or irrelevant features in high-dimensional datasets reduce the classification accuracy of machine learning model, a feature selection algorithm based on approximate Markov blanket is proposed and named as normal max-relevance and min-redundancy (nmRMR) algorithm. Firstly, the algorithm uses the criteria of maximum relevance and minimum redundancy to perform feature relevance ranking. Then, it adopts the approximate Markov blanket to remove redundant features or irrelevant features, and maximize the correlation between features to obtain the optimal feature subset. Experimental results on UCI's eight open datasets show that: the proposed nmRMR algorithm achieves on average 6.875, 20.56 and 3.187 5 reduction in the selected number of feature subsets, as well as 0.78%, 1.88% and 0.825% improvement in the average classification accuracy, compared with the mRMR algorithm, the FullSet algorithm, and the FCBF algorithm, respectively. It is concluded that the proposed nmRMR algorithm is superior to other algorithms.

Keywords: feature selection; feature relevance; redundancy feature; approximate Markov blanket

---

收稿日期：2017-11-21。 作者简介：张俐(1977－)，男，博士生；王枞(通信作者)，女，教授，博士生导师。 基金项目：国家科技基础性工作专项资助项目(2015FY111700-6)。
网络出版时间：2018-08-16 网络出版地址：http://kns.cnki.net/kcms/detail/61.1069.T.20180816.1705.010.html

近年来，基于互信息的特征选择算法受到国内外学者的广泛关注[1]。Zhang 等仅考虑特征与类别标签之间的相关性，而忽视了无关特征或者冗余特征[2]；FCBF 算法[3-4]采用对称不确定性原则作为特征与类别标签之间相关性判断的准则，但是这样所得子集并不是最优子集；mRMR 算法[5-6]采用最大相关最小冗余的准则进行特征子集的排序，使用包装方法进行特征子集的选择，它的问题是特征子集计算时间长，不能有效删除冗余特征。这些不相关和冗余特征往往会降低模型的预测准确性和算法学习效率[7]。因此，非常有必要在数据预处理阶段，对这些不相关和冗余特征进行降维处理[8]。通过降维处理既可以有效降低特征维度，又可以防止模型出现过拟合现象，进而提高算法的精确度。

针对上述问题，本文提出基于近似马尔科夫毯的最大相关最小冗余特征选择(nmRMR)算法，该算法包括特征相关性排序和冗余特征删除两个阶段：第 1 阶段，利用最大相关最小冗余准则进行相关特征排序，并采用前向迭代搜索方法进行最优特征的选择；第 2 阶段，采用近似马尔科夫毯算法删除冗余特征和不相关特征。

1 互信息、mRMR 准则与马尔科夫毯条件

定义 1 设一个随机变量 Y 有 D 个维度，$Y=(y_1,\cdots,y_D)$，根据香农的信息论[9]，一个随机变量 Y 的不确定性可以通过熵 $H(Y)$ 来度量，同理，对于两个随机变量 X 和 Y，它们之间的条件熵 $H(X|Y)$ 表示为当 Y 确定时，对 X 的不确定性的度量。$I(X;Y)$ 是随机变量 X 中包含的关于另一个随机变量 Y 的信息量。因此，$I(X;Y)$、$H(X|Y)$ 和 $H(X)$ 三者之间的关系为

$$I(X;Y)=I(Y;X)=H(X)-H(X\mid Y)=\sum_{y\in Y}\sum_{x\in X}p(x,y)\mathrm{lb}\frac{p(x,y)}{p(x)p(y)} \tag{1}$$

式中：$p(x)$ 和 $p(y)$ 分别为随机变量 X 和 Y 的概率密度，$p(x,y)$ 为联合概率密度。

定义 2 John 认为可以将特征分为强相关特征、弱相关特征和无关特征[9]，因此特征相关性形式化定义如下：F 是特征全集，f_i 是 F 的一个特征，$S_i=F-\{f_i\}$，如果 $I(C|f_i,S_i)\neq I(C|S_i)$，则 f_i 为强相关特征；如果 $I(C|f_i,S_i)=I(C|S_i)$ 且 $\exists S'_i\subset S_i$，并且 $I(C|f_i,S'_i)=I(C|S'_i)$，则 f_i 为弱相关特征；如果 $\exists S'_i\subset S_i$，并且 $I(C|f_i,S'_i)\neq I(C|S'_i)$，则 f_i 为无关特征。

定义 3 特征与标签之间的相关性。假设 S 为特征集合，$|S|$ 为特征数；C 为标签。$I(f_i;C)$ 为特征 f_i 与标签 C 之间的互信息，如果 S 中存在 $1,2,3,\cdots,n$ 个特征，$I(f_i;C)$ 的互信息最大，那么，特征 f_i 表示 S 集合中最强相关特征，依照 mRMR 算法的最大相关准则，是选出满足如下等式的特征的方法

$$\max D(S,C)=\frac{1}{|S|}\sum_{f_i\in S}I(f_i;C) \tag{2}$$

定义 4 特征间的冗余性。假设 F 为特征全集，$S\subset F$，$f_i\in F-S$，$f_j\in S(i\neq j)$，则 $I(f_i;f_j)$ 为特征 f_i 和特征 f_j 之间的互信息。mRMR 算法最小冗余准则满足

$$\min R(S)=\frac{1}{|S|^2}\sum_{f_i\in F-S,f_j\in S}I(f_i;f_j) \tag{3}$$

定义 5 根据 mRMR 算法[7]特征与标签相关性最大、特征间冗余性最小的原则，可得到 mRMR 算法的计算式

$$\max_{f_i\in F-S}\left[I(f_i;C)-\frac{1}{|S|}\sum_{f_j\in S}I(f_i;f_j)\right] \tag{4}$$

定义 6 马尔科夫毯描述[10]。$S\subset F$，$f_i\notin S$，f_i 的马尔科夫毯的条件是 $f_i\perp\{F-S-\{f_i\},C\}\mid S$，其中 $\perp$ 表示独立，$|S$ 表示以 S 为条件。在给定 S 集合中，f_i 独立 $F-S-\{f_i\}$ 和 C，这说明，当 S 集合存在时，f_i 对于标签 C 已经没有任何贡献了，应该立即删除。同时，也说明满足上述条件的最小特征 S 集合为特征 f_i 的马尔科夫毯。

定义 7 近似马尔科夫毯[11]的判断条件。对于特征 f_i 和特征 $f_j(i\neq j)$，特征 f_i 是特征 f_j 的马尔科夫毯条件为

$$\left.\begin{aligned}I(f_i;C)&>I(f_j;C)\\I(f_j;C)&<I(f_i;f_j)\end{aligned}\right\} \tag{5}$$

2 nmRMR 算法描述

本文提出 nmRMR 算法，具体步骤如下：

步骤 1 初始化 F、S 集合并设置 S 集合为空集合；

步骤 2 计算 F 集合中每一个特征与标签之间的互信息，按照互信息大小对 F 集合中的特征进行降序排序；

步骤 3 将 F 集合中的第一个特征 f_i 存入 S 集合，将 f_i 特征从 F 集合中删除；

步骤 4 按照最大相关最小冗余准则进行特征之间的排序操作，将排序好的特征存入 S 集合中；

步骤 5 按照近似马尔科夫毯条件判断，在 S 集合

中进行不相关特征和冗余特征删除操作；

步骤 6　输出最优的特征集合，即 S 集合。

3　实验结果与分析

3.1　nmRMR 算法实验

本文的实验环境是 Intel(R)-Core(TM) i7-500U，内存是 8 GB，Windows7-64 bit 操作系统，仿真软件是 python3.6 和 Jupyter Notebook。数据统计分析软件是 IBM SPSS Statistics20。

实验数据集选择了国际著名的 UCI 通用数据集，数据集涉及医学、工程科学和生物科学等领域，8 个样本数据集包含不同的样本数、特征数和类别数，数据集的具体参数描述见表 1。样本数的范围为 27～4 601，特征数的范围为 23～91，类别数的范围为 2～17。

表 1　实验测试数据集的描述

数据集编号	数据集	样本数	特征数	类别数
1	lung-cancer	27	57	3
2	spect	267	23	2
3	ionosphere	351	35	2
4	dermatology	358	35	6
5	movement_libras	360	91	15
6	wdbc	547	31	2
7	optdigits	562 0	64	17
8	spambase	460 1	58	2

由于某些数据集中具体特征值存在缺失导致样本不完整，在进行实验前，直接删除不完整的样本，根据 Kuargan 提出的类别属性依赖最大化方法[12]离散化处理连续型数据。

本文采用预测分类准确率最为常见的 k-NearestNeighbor 分类器（简称为 KNN）[2] 和 Native Bayes 分类器[13-15]来构建预测模型。

3.2　分类结果实验分析

在实验过程中采用 10 折交叉验证方法进行实验，10 折交叉验证就是每次将数据集平均分成 10 等份，其中 9 份作为训练数据集，1 份作为测试数据集。为了保证公平，每个数据集都做 10 次实验，通过 10 次结果求平均值来求得平均准确率。与 FCBF、mRMR 和 Fullset 算法进行对比，验证本文算法是否可以提高分类的准确率和降低数据的特征数。

比较不同算法优劣性的方式就是两种分类器在 8 个数据集上进行验证。FullSet 算法不进行任何特征排序；mRMR 算法依赖于最大相关最小冗余准则进行特征排序；FCBF 算法和 nmRMR 算法都是在特征初选时对特征进行排序筛选。FullSet 和 mRMR 算法并不具有特征初选的功能。为了保证本实验的公正性，模拟实验都对初选的特征子集不加以限制。

表 2～5 分别给出了 4 种算法在 KNN、Native Bayes 两种分类器中的分类准确率和选择出特征数比较。由表 2 可以看出，在 spambase 数据集中，nmRMR 算法分类准确率略低于 mRMR 算法，但是在其他 7 个数据集上，nmRMR 算法均优于 mRMR 算法。同时，nmRMR 算法的平均分类准确率比 mRMR 算法提高了 0.42%。由表 3 可以看出，nmRMR 算法选择出的平均特征子集数比 mRMR 算法少了 5.375 个。由表 4 可以看出，nmRMR 算法和 mRMR 算法在 lung-cancer 数据集上平均分类

表 2　4 种算法在 KNN 分类器中分类准确率的比较

数据集编号	分类准确率/%			
	nmRMR 算法	FullSet 算法	FCBF 算法	mRMR 算法
1	70.00	58.33	69.98	66.67
2	72.65	69.30	72.65	72.65
3	84.40	84.45	84.40	84.40
4	98.33	96.01	97.76	98.33
5	69.22	67.33	64.33	66.78
6	97.44	97.26	97.26	97.43
7	98.45	98.43	98.52	98.41
8	91.60	91.59	91.87	91.96
平均值	85.00	82.84	84.60	84.58

表 3　4 种算法在 KNN 分类器中选择特征数的比较

数据集编号	选择特征数		
	nmRMR 算法	FCBF 算法	mRMR 算法
1	28	12	23
2	1	1	1
3	4	4	4
4	24	21	21
5	51	87	67
6	16	21	30
7	56	40	63
8	42	44	56
平均值	27.75	28.75	33.125

表 4 4 种算法在 Native Bayes 分类器中分类准确率的比较

数据集编号	分类准确率/%			
	nmRMR 算法	FullSet 算法	FCBF 算法	mRMR 算法
1	70.00	70.00	70.00	70.00
2	70.14	65.21	68.55	68.12
3	69.86	68.7	67.27	68.83
4	98.33	98.00	98.90	98.33
5	54.22	53.11	53.11	53.33
6	78.04	77.86	77.67	76.76
7	90.80	90.39	90.46	90.33
8	90.00	87.89	88.00	89.17
平均值	78.00	76.40	76.75	76.86

准确率相等外，在其他 7 个数据集上，nmRMR 算法均优于 mRMR 算法。同时，nmRMR 算法的平均分类准确率比 mRMR 算法提高了 1.14%。由表 5 可以看出，nmRMR 算法选择出的平均特征子集数比 mRMR 算法少了 8.375 个。由表 2 和表 3 可以看出，nmRMR 算法的平均分类准确率比 FCBF 算法提高了 0.4%，nmRMR 算法选择出的平均特征子集数比 FCBF 算法少了 1 个。由表 4 和表 5 可以看出，nmRMR 算法的平均分类准确率比 FCBF 算法提高了 1.25%，nmRMR 算法选择出的平均特征子集数比 FCBF 算法少了 5.375 个。

表 5 4 种算法在 Native Bayes 分类器中选择特征数的比较

数据集编号	选择特征数		
	nmRMR 算法	FCBF 算法	mRMR 算法
1	14	10	25
2	11	14	12
3	21	23	23
4	21	21	24
5	61	90	89
6	18	25	30
7	56	50	63
8	35	47	38
平均值	29.625	35	38

本实验同时给出 4 种算法在不同分类器上的分类准确率，具体如图 1～图 4 所示。从图 1～图 4 可以看出，nmRMR 算法的分类准确率和特征子集的选择均优于或等于 mRMR、FCBF 和 FullSet 算法。

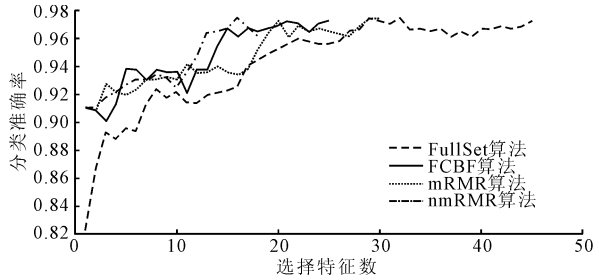

图 1 4 种算法在 KNN 分类器中 wdbc 数据集上的分类准确率

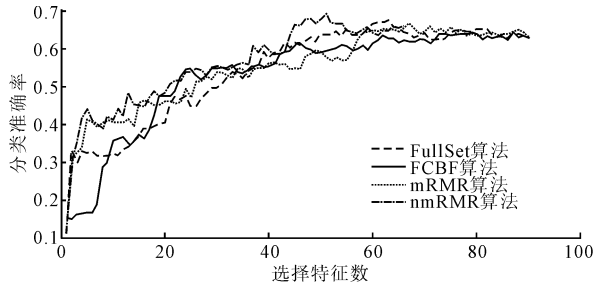

图 2 4 种算法在 KNN 分类器中 movement_libras 数据集上的分类准确率

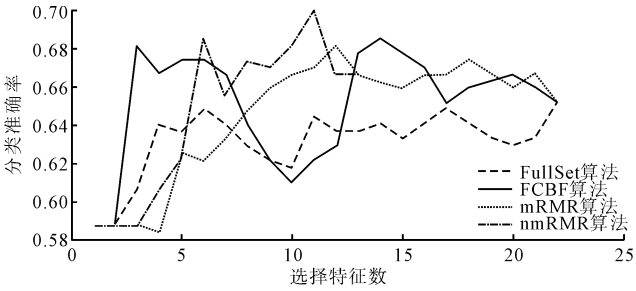

图 3 4 种算法在 Native Bayes 分类器中 spect 数据集上的分类准确率

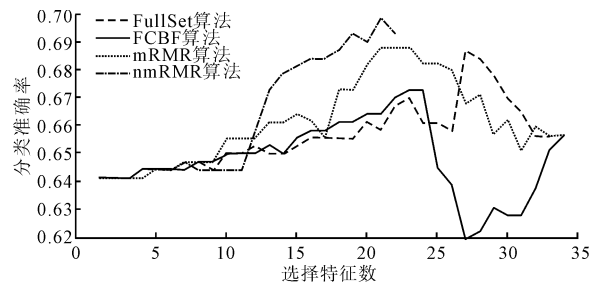

图 4 4 种算法在 Native Bayes 分类器中 ionosphere 数据集上的分类准确率

4 结 论

本文提出了一种基于近似马尔科夫毯的 nmRMR 特征选择算法，该算法在第 1 阶段采用最大相关最小冗余的评价准则进行特征相关性冗余性的分析与排序；第 2 阶段采用近似马尔科夫毯方法

进行特征与标签以及特征间依赖关系的分析，将特征间相互依赖程度高的特征进行删除，保留能够有较高区分能力的特征，组成最优特征子集。由于 nmRMR 算法具有初期特征子集优选能力，它可以进一步地提高分类学习算法的泛化能力。

参考文献：

[1] BENNASAR M, HICKS Y, SETCHI R. Feature selection using joint mutual information maximisation [J]. Expert Systems with Applications, 2015, 42 (22): 8520-8532.

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

[3] LIU Chuan, WANG Wenyong, ZHAO Qiang, et al. A new feature selection method based on a validity index of feature subset [J]. Pattern Recognition Letters, 2017, 92: 1-8.

[4] SUN Xin, LIU Yanheng, LI Jin, et al. Feature evaluation and selection with cooperative game theory [J]. Pattern Recognition, 2012, 45(8): 2992-3002.

[5] PENG H, LONG F, DING C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy [J]. IEEE Transactions Pattern Analysis and Machine Intelligence, 2005, 27(8): 1226-1238.

[6] ZHOU P, HU X, LI P, et al. Online feature selection for high-dimensional class-imbalanced data [J]. Knowledge-Based Systems, 2017, 136(15): 187-199.

[7] GUO Q, ZHANG M. Implement web learning environment based on data mining [J]. Knowledge-Based Systems, 2009, 22(6): 439-442.

[8] WANG Y, WANG J, LIAO H, et al. An efficient semi-supervised representatives feature selection algorithm based on information theory [J]. Pattern Recognition, 2017, 61(1): 511-523.

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

[10] KOLLER D, SAHAMI M. Toward optimal feature selection [C] // Proceedings of the 13th International Conference on International Conference on Machine Learning. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1996: 284-292.

[11] 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.
CUI Zifeng, XU Baowen, ZHANG Weifeng, et al. An approximate Markov blanket feature selection algorithm [J]. Chinese Journal of Computers, 2007, 30 (12): 2074-2081.

[12] KURGAN L A, CIOS K J. CAIM discretization algorithm [J]. IEEE Transactions on Knowledge & Data Engineering, 2004, 16(2): 145-153.

[13] RISH I. An empirical study of the naive Bayes classifier [J]. Journal of Universal Computer Science, 2001, 1(2): 127: 41-46.

[14] BERMEJO P, OSSA L D L, GAMEZ J A, et al. Fast wrapper feature subset selection in high-dimensional datasets by means of filter re-ranking [J]. Knowledge-Based Systems, 2012, 25(1): 35-44.

[15] GOMEZ-VERDEJO V, VERLEYSEN M, FLEURY J. Information-theoretic feature selection for functional data classification [J]. Neurocomputing, 2009, 72 (16/17/18): 3580-3589.

[本刊相关文献链接]

杨宏晖，王芸，孙进才，等. 融合样本选择与特征选择的 AdaBoost 支持向量机集成算法. 2014，48(12)：63-68. [doi：10.7652/xjtuxb201412010]

诸文智，司刚全，张彦斌. 采用邻域决策分辨率的特征选择算法. 2013，47(2)：20-27. [doi：10.7652/xjtuxb201302004]

栗茂林，梁霖，王孙安，等. 结合交叠区异点统计和相关分析的免疫克隆特征选择方法. 2012，46(5)：50-56. [doi：10.7652/xjtuxb201205009]

豆增发，高琳. 利用膜粒子群优化和信息熵的医学文本特征选择. 2012，46(4)：45-51. [doi：10.7652/xjtuxb201204008]

杨宏晖，戴健，孙进才，等. 用于水声目标识别的自适应免疫特征选择算法. 2011，45(12)：28-32. [doi：10.7652/xjtuxb201112006]

薛峰，周亚东，高峰，等. 一种突发性热点话题在线发现与跟踪方法. 2011，45(12)：64-69. [doi：10.7652/xjtuxb201112012]

马超，陈西宏，徐宇亮，等. 广义邻域粗集下的集成特征选择及其选择性集成算法. 2011，45(6)：34-39. [doi：10.7652/xjtuxb201106006]

裴晓梅，郑崇勋. 基于 Fisher 判据时频分析的运动相关脑电特征选择及优化. 2008，42(8)：1026-1030. [doi：10.7652/xjtuxb200808021]

朱虎明，焦李成. 基于免疫记忆克隆的特征选择. 2008，42(6)：679-682. [doi：10.7652/xjtuxb200806007]

崔舒宁，朱丹军，冯博琴，等. 结合受控词汇表的生物基因本体标注与分类. 2008，42(2)：171-174. [doi：10.7652/xjtuxb200802010]

（编辑　武红江）