doi: 10.3969/j.issn.0490-6756.2020.01.013

基于 mRMR 与因子分解机的分类模型研究

王 美, 龙 华, 邵玉斌, 杜庆治
(昆明理工大学信息工程与自动化学院, 昆明 650000)

摘 要: 很多学者用"全球恐怖主义研究数据库"GTD 数据集,采用博弈论、K 近邻法和支持向量机等分析恐怖事件的聚集性,已经取得一些成果.但在前期研究中未有很好考虑数据的稀疏性以及高维度多冗余等会导致聚集分类准确率不高的问题.本文提出一种基于最小冗余最大相关与因子分解机结合的 TFM 分类模型,使用增量搜索方法寻找近似最优的特征解决高维度多冗余问题和 FM 方法解决数据稀疏问题,并对预处理后的恐怖袭击事件数据用 TFM 模型做量化分类.文中使用朴素贝叶斯 NB、支持向量机 SVM、逻辑回归 LR 与 TFM 等 4 个模型的"马修斯相关系数"MCC 进行比较,结果显示 TFM 的 MCC 相对于其他三个模型 NB、SVM、LR 分别提高了 49.9%,2.5%,2.3%,可见 TFM 模型有一定可行性.

关键词: 最小冗余最大相关; GTD; 因子分解机; 马修斯相关系数; TFM 分类模型
中图分类号: TP391 **文献标识码**: A **文章编号**: 0490-6756(2020)01-0096-07

Classification model based on mRMR and factorization machines algorithm

WANG Mei, LONG Hua, SHAO Yu-Bin, DU Qing-Zhi
(Kunming University of Science and Technology, Faculty of Information Engineering and Automation, Kunming 650000, China)

Abstract: Many scholars have made some achievements in aggregation analysis of terrorist events by using the data set of "Global Terrorism Research Database"(GTD) with game theory, k-nearest-neighbor method and support vector machine. However, data sparsity and high-dimensional multi-redundancy are not well considered in the previous research, which may lead to low accuracy of clustering classification. This paper proposes a TFM classification model based on "Minimal-redundancy maximal-relevancy" (mRMR) combined with "Factorization Machines" (FM), in which the incremental search method is used to find approximately optimal features to address the high-dimensional multi-redundancy and the data sparsity is tackled with FM method. TFM model is then used to make quantitative classification on the pre-processed terrorist attack data. The experimental results show the proposed TFM model, in terms of Matthews correlation coefficient (MCC), is increased by 49.9%, 2.5% and 2.3% respectively compared with naive Bayes (NB), support vector machine (SVM) and logistic regression (LR). The comparative result demonstrates that TFM model is feasible to some extent.

Keywords: mRMR; GTD; Factorization machines; MCC; TFM classification model

收稿日期: 2019-04-03
基金项目: 国家自然科学基金(61761025)
作者简介: 王美(1993—), 女, 云南曲靖人, 硕士生, 研究方向为时空数据挖掘.
通讯作者: 龙华. E-mail: longhua@kmust.edu.cn

1 引　言

对恐怖事件聚集性的研究，前期对其研究多是使用对数据汇总统计的方法来显示事件中特定变量随时间变化的趋势. 1999 年 Enders 等人使用时间序列模型，分析恐怖事件对国内生产总值 GDP 的影响，得出当恐怖事件减少时 GDP 呈现增长的结论[1]. 2002 年 Major 使用博弈论与概率分布量化风险，对恐怖事件具体因素隐藏的信息进行分析[2]. 2003 年 Sandler 使用博弈论理性分析了恐怖组织与目标国家间的相互作用，得出一个国家的游客、公民和企业，由于他们威慑能力小，较容易成为攻击目标的结论[3]. Guo 等人在文中分析上述章节所建立的统计与数学模型是基于先验假设和完备的数据集，然而当下因各种原因，GTD 数据集所记录的数据往往不完整或是存在不确定值，所以使得制定有效假设和检验假设能力受限[4]. 故需探索新的可视化和计算方法，以得到分析 GTD 数据更好的模型.

在探索新的分析方法中，首先是 Wang 等人采用协调多视图的方法开发了用于探索恐怖事件数据的可视化分析系统，可揭示一些未知的恐怖袭击事件发生趋势和规律，但该方法中多是对数据汇总统计结果进行可视化，没有对 GTD 数据本身存在的稀疏性以及高维度多冗余问题进行很好的处理[5]. Pagán 发现了该问题，开始引入了数据清洗策略，Pagán 在文中采用均值替换法 MI、中值替换法 MDI 以及 K 近邻填补法 KNNI 对 GTD 中的缺失数据进行填充[6]，清洗后的数据用于线性判别分析 LDA、K 最近邻 KNN 以及分类决策树 RPART 分类器 GTD 中对伊拉克组织的恐怖袭击事件进行分析. 由于 GTD 中存在大量缺失值和冗余特征，全部采用填充方式使得数据集产生了不可忽略的偏差等原因，最终分类效果并不是很理想，交叉验证错误率均在 40%左右. 为了减小 GTD 中的冗余特征量以及合理处理缺失数据，Iqbal 等人[7]对数据进行了降维，并对缺失值增加了删除处理方式，不足的是文中基于主观对问题的理解，对特征采用了手动选择特征方式降维，使得分类结果有很强的主观色彩. 在后面的研究中，处理对 GTD 数据的分类多是采用机器学习方法，2016 年 Muhammad 等人使用监督学习方法如贝叶斯分类器和决策树分类器，对特定巴基斯坦的 GTD 集进行分析[8]；2017 年 Dong 使用 BP 神经网络对印度恐怖袭击事件进行了分析预测[9]；Ding 等人采用经多途径选择的十二个特征，使用神经网络，支持向量机(Support Vector Machine，SVM)和随机森林(RF)宏观预测分析 2015 可能发生恐怖事件的地方[10]. 但就恐怖袭击事件数据本身而言，前面实验中都未有很好的考虑数据集高维度稀疏性大的问题，如高维度稀疏数据会使得 BP 预测效果很差等，针对解决手动特征降维问题，Mo 等人采用了最大相关性(Max-Relevance)以及最大相关性最小冗余 mRMR 特征选择方法获取特征集[11]，并结合支持向量机(SVM)，朴素贝叶斯(Naive Bayes，NB)和 Logistic 回归(LR)分类器对恐怖袭击事件进行分类研究，文中有效地验证了将机器学习应用于恐怖主义研究领域的可行性，但文中训练集是采用专家定义对事件进行分类，且未解决选取后的优化特征集仍存在的数据稀疏性的问题，这将会影响分类效果不佳.

基于以上问题，同时与文献[12]提出的混合变量选择算法相比，因 FM 往往有很好的通用性能，且常能取得较好的效果且善于处理稀疏数据集的特性[13]，故本文提出一种基于“最小冗余最大相关”(Minimal-redundancy maximal-relevancy，mRMR)与“因子分解机”(Factorization Machines，FM)结合的 TFM 分类模型，mRMR 使用增量搜索方法寻找近似最优的特征集，找到最佳特征子集以最大化分类精度，并结合因子分解机 FM 模型对 GTD 进行量化分类.

2 方法与模型应用

2.1 数据预处理

公开数据集“全球恐怖主义研究数据库”(Global Terrorism Research Database，GTD)[14]具有高维海量数据，大数据的低价值的特征，存在大量未记录缺失数据，为了提高数据的质量，所以需要进行数据预处理.

文中用二分类法对 GTD 展开讨论，假设 GTD 数据集表示为

$$D=\{x^{(1)},\ x^{(2)},\ \cdots,\ x^{(i)}\},\text{则 } x^{(i)}=\begin{bmatrix}x_1^{(i)}\\x_2^{(i)}\\\vdots\\x_j^{(i)}\end{bmatrix}\in R^j,\ i\in(1,m),\ j\in(1,n)$$

其中，m 表示总的事件数；n 表示 GTD 数据集总的维度.

定义 1 定义 GTD 事件类型为 $y^{(i)}=\{0,1\}$，若此处以 1 表示 A 组织所为事件，则 0 与之相反，则测试集表示为$(x^{(1)},y^{(1)}),\cdots,(x^{(i)},y^{(i)}),i\in(1,m)$，$m$ 表示总的事件数，因 GTD 数据具有高损失率的问题，采用数据剔除法进行预处理.

定义 2 定义特征 j 的缺失率为

$$p_j=\frac{\sum_{i=1}^{m}\text{missing}(x_j^{(i)})}{m} \tag{1}$$

其中，missing( ) 函数统计缺失值个数；p_j 表示第 j 个特征的缺失率；m 表示总的事件数量. 对第 k 个特征，若 $p_k>\delta$(为某阈值)时，认为其缺失值过大，将其剔除，此时预处理后的数据 $x'(i)=\{x_1^{(i)},x_2^{(i)},\cdots,x_j^{(i)},\cdots,x_L^{(i)}\}$，$L<n$ ，$x_L^{(i)}$ 表示第 i 个事件的第 L 个特征值.

2.2 mRMR 特征选择

mRMR 算法[15]是根据特征与目标分量间相关性以及特征与特征间的冗余性进行特征提取，其中，相关性和冗余性用互信息量(MI)值大小表示；MI 与相关性成正比，互信息的定义如下.

假设预处理后的 GTD 数据集为 $D'=\{x'^{(1)},x'^{(2)},\cdots,x'^{(m)}\}$，其中第 i 个特征表示为 $x_i=\{x_i^{(1)},x_i^{(2)},\cdots,x_i^{(m)}\}$，$i\leqslant n$ ，那么第 i 个特征与第 j 个特征的互信息量为

$$MI(x_i;x_j)=\sum_{x_i^{(m_1)}\in x_i}\sum_{x_j^{(m_2)}\in x_j}p(x_i^{(m_1)},x_j^{(m_2)})\log\frac{p(x_i^{(m_1)},x_j^{(m_2)})}{p(x_i^{(m_1)})p(x_j^{(m_2)})},m_1,m_2\in(1,m) \tag{2}$$

在式(2)中，$p(x_i^{(m_1)},x_j^{(m_2)})$ 为联合概率密度；$p(x_i^{(m_1)})$和 $p(x_j^{(m_2)})$为边缘概率. 设 G 为包含 D' 中的所有特征向量的集合；G_s 表示已经经过 mRMR 筛选的 s 个特征向量；G_t 表示待选择的 t 个特征向量集. 那么类别变量 c 与 G 内的特征量 x_f 相关性 O 计算如下式.

$$O=MI(x_f;c),\ f\leqslant n \tag{3}$$

G_t 中的特征向量 x_t 与 G_s 中的全部特征向量间的冗余 R 度计算如下式.

$$R=\frac{1}{s}\sum_{x_s\in G_s}MI(x_t;x_s),\ t<n,s<n \tag{4}$$

假设 G_s 已有 p 个特征，则 G_s 为了从 G_t 中得到满足最大相关和最小冗余的第 $(p+1)$ 个特征向 $x'_{(p+1)}$，结合式(3)和式(4) 得到的如下式(5).

$$x'_{(p+1)}=\max_{x_t\in G_t}[MI(x_t;c)-\frac{1}{m}\sum_{x_s\in G_s}MI(x_t;x_s)] \tag{5}$$

当总的特征为 $N(=s+t)$时，mRMR 的特征评估将进行 N 轮. 经过 mRMR 评估后原特征集 $\{x_1,x_2,\cdots,x_N\}$，重新排序后的特征集表示为 s 如下.

$$s=\{x'_1,x'_2,\cdots,x'_N\}$$

最终最优特征集取目标函数最小损失值对应的 s 中的 $H(\leqslant N)$个特征.

2.3 TFM 分类算法

TFM 算法意在构建分类模型，结合了因子分解机算法 FM[16]和分类阈值算法.

FM 预测算法通过分解交互参数来模拟变量对与目标间的所有交互量，假设取事件 o 表示为 $x^{(o)}=\{x_1^{(o)},x_2^{(o)},\cdots,x_n^{(o)}\}$，则一个二阶的 FM 算法输出值 $\hat{y}$ 定义如下.

$$\hat{y}(x^{(o)})=w_0+\sum_{i=1}^{n}w_ix_i^{(o)}+\sum_{i=1}^{n}\sum_{j=i+1}^{n}<v_i,v_j>x_i^{(o)}x_j^{(o)} \tag{6}$$

式(6)中，$x_i^{(o)}$ 为每个维度的特征分量；v_i 为引入辅助向量$(=(v_{i1},v_{i2},\cdots,v_{ik})^T\in R^k,\ i\in(1,n)$；其中，$k\in N+$为超参数)，$<v_i,v_j>$表示两个大小为 k 的向量 v_i 和向量 v_j 的点积，模型中表示第 i 个变量与第 j 个变量的交互参数，w_0 表示全局偏值，$w_o\in R$；w_i 表示第 i 个特征变量的影响程度，$W\in R^n$.

式(6) 计算复杂度为 $O(kn^2)$，根据文献[16]将式(6)计算复杂度化简为 $O(kn)$后，表达式如下式.

$$\hat{y}(x^{(o)})=w_0+\sum_{i=1}^{n}w_ix_i^{(o)}+\frac{1}{2}\sum_{i=1}^{k}((\sum_{j=1}^{n}v_{ij}x_i^{(o)})^2-(\sum_{i=1}^{n}v_i^2x_i^{(o)2})) \tag{7}$$

假设 FM 中所有参数 $\Theta=\{w_0,w_1,\cdots,w_n,v_{11},\cdots,v_{kn}\}$，为了避免产生过拟合现象，所以此处 FM 最优化目标选取正则化的最小二乘法.

$$\Theta^*=\underset{\Theta}{\text{argmin}}\sum_{i=1}^{N}(\text{loss}(\hat{y}(x^{(i)}),y^{(i)})+\sum_{\theta\in\Theta}\lambda_\theta\theta^2) \tag{8}$$

其中，$x^{(i)}$表示第 i 个事件；loss( )为损失函数；λ_θ 表示参数 θ 的正则化系数. 结合以上分析，模型中参数$\{w_0,W,V\}$ 可采用随机梯度下降法 SGD 对进行学习，计算伪代码可见 ALGORITHM 1.

分类阈值算法，加入该算法以实现二分类，假

设训练集中属于 A 分类的有 z 件，于 A 分类的有 a 件，则取阈值函数：

$$p'=\frac{z}{z+a} \tag{9}$$

当事件 $x^{(i)}$ 预测值 $P>P'$ 时，预测结果属于 A 分类，反之不属于.

TFM 模型是将 FM 计算得到的预测结果，与分类阈值结果比较然后归类得到分类结果.

ALGORITHM 1：随机梯度下降法(SGD)

Input：训练集 S，正则化系数集 λ，学习率 η，正态分布方差参数 σ
Output：模型参数 $\Theta=(w_0,w,V)$
Initilization：$\omega_0:=0;w:=0;V\sim N(0,\sigma)$;
Repeat

For $(X,y)\in S$ Do
$\omega_0=\omega_0-\eta\left(\frac{\partial}{\partial\omega_0}\text{loss}(\hat{y}(x),y)+2\lambda^0\omega_0\right)$;
For $i\in\{1,2,\cdots,n\}x_i\neq 0$ Do
$\omega_i=\omega_i-\eta\left(\frac{\partial}{\partial\omega_i}l\text{oss}(\hat{y}(x),y)+2\lambda\omega_{\pi(i)}\omega_i\right)$;
For $j\in\{1,\cdots,k\}$ do
$v_{ij}=v_{ij}-\eta\left(\frac{\partial}{\partial\omega_i}\text{loss}(\hat{y}(x),y)+2\lambda v_{\pi<i>,j}v_{ij}\right)$;
end
end
end
untilstopping criterion is met;

2.4　模型评估

2.4.1　logit loss 损失函数　logit loss 是普遍认可的一种模型分类效果分类标准[17]，其定义如下.

$$\text{loss}(\hat{y},y)=-ln\sigma(\hat{y}y) \tag{10}$$

其中，$\sigma(x)=\frac{1}{1+e^{-x}}$ 为 sigmoid 函数；y 为真实值；为预测值，当 $\hat{y}$ 和 y 越接近时，损失 loss$(\hat{y},y)$就越小. 文中根据 logit loss 的最小损失函数值确定选取最优特征集.

2.4.2　分类指标　分类模型性能评价指标中常使用的有准确率、召回率和灵敏度[18]，以及本文采用的马修斯相关系数(Matthews Correlation Coefficient，MCC)[19]衡量指标：

(1) 准确率：Accuracy$=\frac{TP+TN}{TP+TN+FP+FN}$，预测正确的数占样本数的比例；

(2) 灵敏度或召回率：Recall$=\frac{TP}{TP+FN}$，阳性值中实际被预测正确所占的比例；

(3) 特异度：$SPC=\frac{TN}{FP+TN}$，阴性值中实现被预测正确所占的比例；

(4) 马修斯相关系数：$MCC=\frac{TP\times TN-FP\times FN}{\sqrt{(FP+TN)(TP+FN)(TN+FP)(TN+FN)}}$

上式中的 TN、FN、TP 以及 FP 定义如表 1 所示.

表 1　分类评价指标定义

Tab. 1　Classification evaluation indicator definition

实际类别	预测类别	
	0	1
0	True Negative(TN)	False Positive (FP)
1	False Negative(FN)	True Positive (TP)

TN 为被模型预测为负的负样本；FN 为被模型预测为负的正样本；TP 为被模型预测为正的正样本；FP 为被模型预测为正的负样本.

文中使用了不同的模型进行分类效果比较，鉴于 MCC 同时考虑了 TP、TN、FP 和 FN，可用来很好的衡量分类效果，MCC 越大越好，故本文选用 MCC 评价指标.

3　实验及结果

实验中，我们采用了全球恐怖主义研究数据库 GTD 数据集. GTD 记录了从 1970 年至今世界各地的恐怖事件信息，并且不断的更新各种恐怖事件，至今已超过 14 万恐怖袭击事件，且每一个事件超过 45 个特征记录值，这使其成为目前是介绍基于恐怖事件的最全面的非机密数据. 文章中选取了 2001 年至 2017 年近 17 年南亚地区 A 组织所为事件分别打标签得到分析数据集 G. 表 2 列出分析对象近 17 年的 18 737 个总事件，其中分别对 A 组织所为和非其所为事件数进行统计，并打标签，A 组织所为标记 1，反之标记 0.

表 2　恐怖事件组织和标签

Tab. 2　Terrorist organization and labeling

Type	Data	
	Sample	Lable
A 组织	7 473	1
非 A 组织	11 264	0

我们将按照 2.1 方法对数据集 G 进行处理，处理后的数据集 G 中 $p_j>0$ 有一半特征，说明处理

后的 GTD 数据集中仍存在大量未记录数据或空值数据.

3.1 获取最优特征子集

我们于 Python 3.6 上完成实验,实验中首先使用 mRMR 函数应用于 A 组织训练集,根据式(5)增量特征选择方法获取诸多特征子集,实验有 58 个特征故共有 58 个候选子集.我们对训练集使用四折交叉检验获取训练集和验证集,图 1 所示为每一个候选子集对应的 FM 预测输出与分类结果的损失值,我们从图中观察与后台计算结果对应取最小损失值 6.044 257 3 对应的 36 个特征子集为最优特征集.

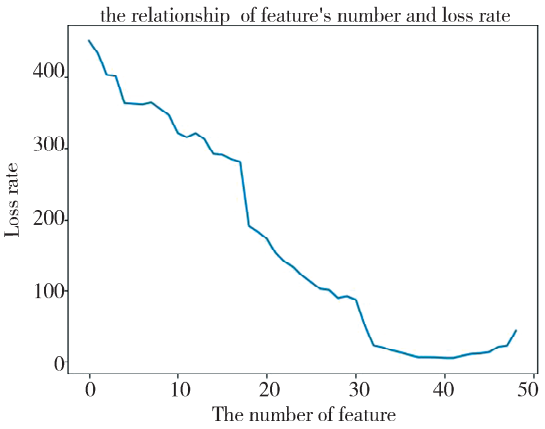

图 1 mRMR 选取特征子集对应的分类损失值

Fig. 1 Loss value for feature subset based on mRMR

3.2 FM 与其他模型比较

许多方法都用于分类模型中,如文中提到的逻辑回归(Logistic Regression, LR)、支持向量机 SVM 和朴素贝叶斯,但是很难说明哪种分类比较好,不同的数据集适合不同的分类模型.本文基于 mRMR 确定的最优特征集应用 GTD 数据集对 TFM、LR、SVM 和 NB 其分类进行了实验,实验中采用了 tensorflow 框架实现 TFM 模型,极大化的优化了模型的实现,分类具体效果比较见表 3.表 3 中耗时为测试时长 T,即测试所用时间.由表 3 可直观看出 TFM 分类准确率与 MCC 值均优于三个模型,LR 次之,NB 准确率和 MCC 最为不理想,但就耗时来说 TFM 相对微大于三个模型,就此问题来分析,分类结果供人为使用分析,故毫秒的差异不足以影响使用效果.结合结果与 LR[20]、SVM[21]以及 NB[22]模型特点可分析,TFM 将特征映射到高维来考虑特征与特征间的关系,且 TFM 通过引入辅助向量迭代求取两个特征间的参数,故效果较好,但也正是引入了辅助向量,故相对于其他三个模型来说耗时略大些.

从表 3 可知,TFM 分类效果较稳定,准确率也较高.表 4 所示事件即为 TFM 分类正确,但其他三个模型失效的案例.C_1 是 2006 年 11 月 19 日的一个暴恐事件,属于 A 组织所为,正确归类应是 1;C_2 是 2008 年 7 月 7 日的一个暴恐事件,非 A 组织所为,正确归类应是 0.TFM 的分类错误率 1.7% 低于 LR 的 2.95%、SVM 的 3.03% 以及 NB 的 26.56%,所以在 GTD 分类模型中可优先考虑 TFM 模型.

表 3 TFM 与其他模型比较

Tab. 3 TFM compared with LR, SVM and NB

模 型	准确率 Accuracy/%	马修斯相关系 MCC	测试 T/s
TFM	98.23	0.963 30	0.374 818
LR	97.05	0.940 32	0.017 136
SVM	96.97	0.938 66	0.252 786
NB	73.44	0.463 85	0.021 237

表 4 事件 C 分类结果

Tab. 4 The classification result of event C

事件	经度	纬度	lable	分类结果			
				LR	SVM	NB	TFM
C_1	70.027 089	32.874 079	1	0	0	0	1
C_2	69.147 011	34.516 895	0	1	1	1	0

4 结果与讨论

从实验可知,TFM 模型有一定优势,但存在耗时 T 略大问题,分析是因为 TFM 模型中引入辅助向量导致.在以下场景中该问题可做如下优化:(1) 如类似本文分类结果供人为使用分析的分类中,毫秒的差异不足以影响使用效果可接受;(2) 选择适合的数据集,因为该模型原本就为解决稀疏数据分类问题而提出,故合适的数据集,即使耗时稍大,但效果明显,如此处采用一个稀疏度约为 0.029 6 的 0 和 1 组成的数据集(此处定义:稀疏度=非零总数值个数 / 总数值个数)做个测试实验,结果如表 5 所示,值得分析的是从表 5 中看 LR 与 SVM 准确率相同,侧面反映出 SVM 将特征映射到高维部分计算失效,只有线性部分有用,正如文献[16]说的处理稀疏数据时 FM 比 SVM 更有优势,当数据过于稀疏时 SVM 失效,所以最终结果

与 LR 一致，同时说明在处理稀疏数据集时 FM 有利于高维部分的计算；(3) 类似医学中癌症症状分类，或关键时刻的模糊稿件字符识别分类如罪犯笔迹等中，记录数据较为稀疏且更为要求分类性能稳定，效果较好的情况，可以一定的时间为代价，得到更准确的分类效果，此时可考虑 TFM 分类模型.

表 5　运用 TFM 模型稀疏测试数据结果

Tab. 5　Sparse-data test results with the TFM model

模 型	准确率 Accuracy/%	耗时 T/s
TFM	100	0.002
LR	90.48	0.001
SVM	90.48	0.001

5　结　论

由以上分析可知，类似于 TFM 和 LR 类的机器学习方法可用于分析恐怖主义数据的特征关系，且具有高准确性和快速性. 本文基于 GTD 数据集，使用 Python3.6 统计数据，使用 mRMR 算法获取分类损失函数值最小的优选特征子集，对准备好的数据集使用不同的分类器进行分析. 我们得出结论：TFM 优于其他三个模型，但其代价是耗时略多，但也有一定的优化方案在第四部分中进行了概括. 所以总体来说分类 GTD 时，若需要更为准确的分类效果则选 TFM，若更为追求时间成本，则可选用 LR 或 SVM 算法. 当然在实验中遇到特别稀疏的数据集，TFM 模型较合适. 总的来说考虑了 mRMR 特征提取算法，结合了 FM 模型与分类阈值算法的 TFM 模型确实量化了对恐怖事件的分类问题，之后可在此基础上进行更多的研究. 如文献[23]中提出的改进 mRMR 特征选择方法，融合多种相关度量方式对特征子集进行特征选择，可以恐怖主义事件为数据集展开优化 mRMR 算法问题以及考虑多分类问题等.

参考文献：

[1] Enders W, Sandler T. Transnational terrorism in the post-Cold War era [J]. Int Stud Quart, 1999, 43: 145.

[2] Major J A. Advanced techniques for modeling terrorism risk [J]. J Ris Financ, 2002, 4: 15.

[3] Sandler T. Terrorism & game theory [J]. Simul Gaming, 2003, 34: 319.

[4] Guo D, Liao K, Morgan M. Visualizing patterns in a global terrorism incident database [J]. Environ Plann B: Plann Des, 2007, 34: 767.

[5] Wang X, Miller E, Smarick K, *et al*. Investigative visual analysis of global terrorism [C]//Computer Graphics Forum. Oxford, UK: Blackwell Publishing Ltd, 2008.

[6] Pagán J V. Improving the classification of terrorist attacks a study on data pre-processing for mining the Global Terrorism Database[C]// International Conference on Software Technology & Engineering. [s. l.]: IEEE, 2010.

[7] Iqbal R, Murad M A A, Mustapha A, *et al*. An experimental study of classification algorithms for crime prediction [J]. Indian J Sci Techn, 2013, 6: 4219.

[8] Muhammad H, Kazi H. Use of predictive modeling for prediction of future terrorist attacks in pakistan [J]. Int J Comput Appl, 2016, 179: 8.

[9] Dong Q L. Machine learning and conflict prediction: A cross-disciplinaryapproach [J]. World Econ Polit, 2017, 7: 100.

[10] Ding F, Ge Q, Jiang D, *et al*. Understanding the dynamics of terrorism events with multiple-discipline datasets and machine learning approach [J]. Plo Sone, 2017, 12: e0179057.

[11] Mo H, Meng X, Li J, *et al*. Terrorist event prediction based on revealing data[C]//2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA). [s. l.]: IEEE, 2017.

[12] 赵伟卫，李艳颖，赵风芹，等. 基于互信息和随机森林的混合变量选择算法[J]. 吉林大学学报：理学版，2017，55：933.

[13] 燕彩蓉，周灵杰，张青龙，等. 因子分解机模型的宽度和深度扩展研究[J]. 软件学报，2019，30：822.

[14] LaFree G, Dugan L. Introducing the global terrorism database [J]. Terrorism Political Violence, 2007, 19: 181.

[15] Peng H, Long F, Ding C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy [J]. IEEE T Pattern Anal, 2005, 27: 1226.

[16] Rendle S. Factorization machines[C]//2010 IEEE International Conference on Data Mining. [s. l.]: IEEE, 2010.

[17] Zou H, Zhu J, Hastie T. The margin vector, admissible loss and multi-classmarginbased classifiers [J]. Ann Appl Statist, 2006, 2: 1290.

[18] 秦锋，杨波，程泽凯. 分类器性能评价标准研究[J]. 计算机技术与发展，2006，16：85.

[19] Smialowski P，Martin-Galiano A J，Mikolajka A，*et al*. Protein solubility：sequence based prediction and experimental verification[J]. Bioinf，2007，23：2536.

[20] Arabameri A，Pradhan B，Rezaei K，*et al*. Spatial modelling of gully erosion using evidential belief function，logistic regression，and a new ensemble of evidential belief function logistic regression algorithm[J]. Land Degrad Dev，2018，29：4035.

[21] 秦璐，李旭伟. 基于区域标记法的代价敏感支持向量机在股票预测中的研究[J]. 四川大学学报：自然科学版，2018，55：277.

[22] Jiang L，Zhang L，Li C，*et al*. A correlation-based feature weighting filter for Naive Bayes[J]. IEEE T Knowl Data En，2019，31：201.

[23] 王华华，黄龙，周远文，等. 改进的 mRmR 特征选择方法在人体行为识别中的应用[J]. 重庆邮电大学学报：自然科学版，2019，31：261.

引用本文格式：

中　文：王美，龙华，邵玉斌，等. 基于 mRMR 与因子分解机的分类模型研究[J]. 四川大学学报：自然科学版，2020，57：96.

英　文：Wang M，Long H，Shao Y B，*et al*. Classification model based on mRMR and factorization machines algorithm[J]. J Sichuan Univ：Nat Sci Ed，2020，57：96.