基于全球恐怖主义数据库的特征选择方法研究

姜国庆[1] 赵 梦[2] 杨 涛[1] 彭如香[1] 孔华锋[3*]

[1](公安部第三研究所 上海 201204)

[2](西安电子科技大学 陕西 西安 710126)

[3](武汉商学院 湖北 武汉 430056)

摘 要 恐怖主义被称为现代人类社会之癌,是世界各国政府和人民面临着的重大的挑战,应该引起全人类的重视。在使用全球恐怖主义数据库中的数据对恐怖主义活动进行研究时,从高维数据中提取关键的特征,是反恐研究中的重点和难点。针对全球恐怖主义数据库中特征的高维性、冗余性和数据不完整性的特点,分别采用最小冗余最大相关算法(mRMR)、基于支持向量机的递归删除算法(SVM-RFE)和基于随机森林的特征选择算法进行特征筛选与提取。利用 K-近邻(KNN)分类器其对上述特征选择方法进行降维结果分析和分类结果比较。实验结果表明,特征选择算法不仅能提高分类性能还能提高分类效率,并且基于支持向量机的递归删除算法(SVM-RFE)选择的特征子集在预测恐怖主义活动时准确率更高。

关键词 全球恐怖主义数据库 特征选择 mRMR SVM-RFE 随机森林

中图分类号 TP3 **文献标识码** A DOI:10.3969/j.issn.1000-386x.2019.04.006

FEATURE SELECTION METHOD BASED ON GLOBAL TERRORISM DATABASE

Jiang Guoqing[1] Zhao Meng[2] Yang Tao[1] Peng Ruxiang[1] Kong Huafeng[3*]

[1](*Third Research Institute of Ministry of Public Security, Shanghai 201204, China*)

[2](*Xidian University, Xi'an 710126, Shaanxi, China*)

[3](*Wuhan Business University, Wuhan 430056, Hubei, China*)

Abstract Terrorism, known as the cancer of modern human society, is a major challenge faced by governments and people around the world and should be brought to the attention of all mankind. When using the data in the global terrorism database to study terrorist activities, extracting key features from high-dimensional data is the focus and difficulty in counter-terrorism research. According to the characteristics of high dimensionality, redundancy and data incompleteness in the global terrorism database, we adopted minimum-redundancy maximum-relevancy (mRMR), recursive feature elimination based on support vector machine (SVM-RFE) and the feature selection algorithm based on random forest respectively to screen and extract features. K-nearest neighbor (KNN) classifier was used to analyze the dimension reduction results and compare the classification results of the above feature selection methods. The experimental results show that the feature selection algorithm can improve not only the classification performance but also the classification efficiency. And the feature subset selected by SVM-RFE has higher precision when predicting terrorist activities.

Keywords Global terrorism database Feature selection mRMR SVM-RFE Random forest

---

收稿日期:2018-10-22。公安部科技强警基础工作专项项目(2018GBJC19);江西省经济犯罪侦查与防控技术协同创新中心开放基金资助课题项目(JXJZXTCX-024)。**姜国庆**,助理研究员,主研领域:信息安全,数据挖掘。**赵梦**,硕士。**杨涛**,副研究员。**彭如香**,助理研究员。**孔华锋**,研究员。

0 引 言

近些年来,恐怖主义活动也越来越活跃。就今年而言,无论是新疆鄯善“6·26”严重暴力砍人事件,还是加拿大多伦多“7·22”恶性枪击案件,都表明了恐怖主义袭击给人类社会的政治、经济等各方面造成的不良影响。对此,应积极采取预防措施以避免恐怖主义事件的发生或是减少恐怖主义带来的损失。

随着信息技术的发展,将信息技术应用到反恐怖主义的研究中[1],将有利于遏制恐怖主义的发展。国外有些国家利用信息技术现在已经建立了比较完善的反恐怖主义技术体制,将反恐怖主义研究和人工智能相结合,利用恐怖主义数据来预测恐怖主义组织及其行为。中国在反恐怖主义研究是从 21 世纪开始的,主要偏向于反恐基础理论、立法保障等方面,很少有将人工智能和恐怖主义数据结合来进行研究的。目前研究反恐的恐怖主义数据集主要是全球恐怖主义数据库,数据库中记录的每一个恐怖主义事件的变量超过 40 个。特征选择是从一组特征中挑选出一些最有效的特征以降低特征空间维数的过程[2]。特征选择是模式识别的关键步骤之一,它不仅能够减少特征数量、避免维数灾难,还能提高预测模型的泛化能力。将不同特征选择方法应用到恐怖主义预测模型中,将得到一个有效的特征选择方法来提高预测模型的性能。

1 全球恐怖主义数据库

全球恐怖主义数据库是一个开源的数据库,它记录了从 1970 年至今的全球恐怖主义事件的信息,现由美国恐怖主义研究和应对全国联盟(START)支持和管理。全球恐怖主义数据库也是一个动态的数据库,每一年都会更新一次,是目前研究恐怖主义最全面的公开的数据源。其特点如下:

(1) 高维性。全球恐怖主义数据库共含有 135 个特征,分别记录了恐怖主义事件的九大类信息:GTD 的标志号和日期、事件信息、事件发生的地点、攻击信息、武器信息、目标/受害者信息、凶手信息、伤亡和后果、附加信息和来源。

(2) 冗余性。在这九类中,GTD 的标志号和信息中特征个数最少为 7,伤亡和后果特征个数最多为 29,每一类中的特征表示含义都十分相似,冗余度很高。

(3) 数据量大。截至 2016 年 6 月,全球恐怖主义数据库中共记录了超过 17 万件恐怖主义事件,平均每年增长 3 700 多条记录,近几年来更是以每年上万条的速度在增长,数据量一直在增加。

(4) 数据不完整性。全球恐怖主义数据库中虽然有 135 个特征,超过 17 万条数据,但是受到恐怖主义事件的不确定性、未知性和收录的来源性的影响,很多特征的特征值都是缺失的。据统计,数据库中所有记录的 135 个属性中完全不存在缺失值的记录为零。

正是由于上述全球恐怖主义数据库以上的特点,在使用其直接预测恐怖主义活动时是不可行的。而如何快速又高效地处理这些数据,成为反恐预测中的重要环节,但是国内外学者对此方面的研究少之又少。莫豪文[3]利用最小冗余最大相关特征选择算法选择的特征子集来预测恐怖主义袭击类型,并取得了较好的效果。Mo 等[4]介绍了最大相关性和最小冗余最大相关性两种特征选择算法,并将两种算法应用到在预测恐怖主义组织时的特征降维。其他出现的对于全球恐怖主以数据库中的特征处理都多以人工选择为主[5-6]。以上学者的研究给出了全球恐怖主义数据库特征选择的方向。

2 特征选择方法

从实现方式上讲,机器学习特征的选择方法可以归纳为以下几类:1) 过滤式;2) 包裹式;3) 嵌入式。过滤式主要是通过单个特征和结果之间的关联,例如相关系数、卡方检验、信息增益和互信息来决定特征子集的选取。包裹式依赖机器学习算法,通过学习算法决定每一次增加哪一个特征。嵌入式则是直接使用机器学习中的学习器,对学习器训练并自动选择特征。本文分别选取了这三类中的较为典型的特征选择算法进行介绍和实验。

2.1 最小冗余最大相关性算法

最小冗余最大相关算法(mRMR)是一种过滤式的特征选择算法。它是一种基于两个变量之间的互信息的优选方法,互信息[7]表示两个随机变量之间的相关性[8],它是信息论中的一个概念。最小冗余最大相关算法在特征选择中主要是在特征集合中找出特征向量和类别向量中互信息最大的,同时特征向量之间的互信息最小的特征向量。

其中,最大相关性:

$$\max D(S,c) \quad D = \frac{1}{|S|}\sum_{x_i \in S} I(x_i;c) \tag{1}$$

式中:S 为特征集合;c 为类别向量;x_i 为第 i 个特征。

最小冗余度:

$$\min R(S) \quad R = \frac{1}{|S|^2}\sum_{x_i,x_j \in S} I(x_i,x_j) \tag{2}$$

所以,mRMR 特征选择的标准为:

$$\min\phi(D,R) \quad \phi = D - R \tag{3}$$

2.2 基于支持向量机的递归删除算法

基于支持向量机的递归删除算法(SVM-RFE)是一种包裹式的特征选择算法,也是一种基于 SVM 中最大间隔原理的后向序列约简算法[9]。SVM-RFE 在特征选择选择时,主要是先利用支持向量机模型训练数据集,得到每一个特征的权重并将权重作为得分标准进行排序;接着在使用后向序列选择算法移除得分最小的特征;然后对剩下得分高的特征的数据集再使用支持向量机训练;重复上述过程,直至没有可移除的特征,最终得到一个特征的重要性排序表。具体算法如下:

1) 输入:训练样本 $X = [x_1, x_2, \cdots, x_n]$ 和类别标签 $y = [y_1, y_2, \cdots, y_n]$。

2) 初始化:特征集合 $S = [1, 2, \cdots, k]$,特征排序 $r = [\cdot]$。

3) 特征排序过程:

a) 循环下列过程直至 $S = [\cdot]$

b) 获取当前训练样本 $X = [x_1, x_2, \cdots, x_n]$

c) 给定参数,训练 SVM 分类器,计算每个特征的权重 w

d) 对特征权重进行排序,并找出权重值最小的特征 $f = \min(x)$

e) 更新特征排序表 $r = [S(f), r]$,移除权重最小的特征

4) 输出:特征排序表 r。

2.3 基于随机森林的特征选择算法

基于随机森林的特征选择算法是一种嵌入式的特征选择算法。随机森林[10-11]是一种集成学习器,它由一组决策树[10]分类器构成。随机森林由于其鲁棒性好和准确率高的特点,分类能力也很强。在使用随机森林进行分类时,主要分为两步:一是使用 Bagging 将训练集分为若干个训练子集;二是在每个训练子集上分别建立一个决策树,对相应的训练子集进行分类,最终分类结果由每一个决策树投票而得。在使用随机森林进行特征选择时,也用到了分类中的过程,但是结果并不是分类结果,而是得到特征的重要性评估。此方法主要分为两个步骤:单特征重要性评估和特征剪除。

单特征重要性评估步骤:

1) 计算随机森林中 N 棵决策树的袋外误差,记作 OOB_{Error1}。

2) 随机对袋外数据中的所有特征进行加躁处理,并计算此时的袋外误差 OOB_{Error2}。

3) 单特征的重要性为 $\sum (OOB_{\text{Error2}} - OOB_{\text{Error1}})/N$。

在本文中根据单特征重要性,使用序列后向选择算法进行特征剪除。

3 实验及结果分析

3.1 实验设计

由于全球恐怖主义数据库的特征冗余性和数据不完整性的特点,直接对 135 个特征进性特征选择会出现数据集特征值不完整的情况。为了拥有完整的数据集,本文选取了特征值相对完整的 46 个特征进行特征选择。此外,gname 属性将作为类别标签,用于分类预测。

在实验中,本文以预测恐怖主义组织为目标,主要是预测全球恐怖主义数据库中活动频繁的前 30 个恐怖主义组织,故以恐怖主义组织这一特征为目标变量,对 46 个特征进行特征选择。实验借助 Python 3.0 中的 sklearn 库,采用了三种特征选择算法:最小冗余最大相关算法(mRMR)、基于支持向量机的递归删除算法(SVM-RFE)和基于随机森林的特征选择算法。之后对三种特征选择算法选择出来的特征使用 K-近邻为学习模型对 30 个恐怖主义组织进行分类验证,分析比较特征选择算法的性能。

对于恐怖主义组织的分类预测中,使用准确率作为分类性能的衡量标准。准确率的定义为:

$$Acc = \frac{\sum_j TP_j + \sum_j TN_j}{\sum_j |S_j|} \tag{4}$$

式中:$\sum_j TP_j$ 表示第 j 个恐怖主义组织被正确分类的个数;$\sum_j TN_j$ 表示第 j 个恐怖主义组织被错误分类的个数;$\sum_j |S_j|$ 表示第 j 个恐怖主义组织的样本总数。

3.2 实验结果及分析

图 1 是使用三种特征选择算法得出的这 46 个特征的重要性排序,然后使用前向序列选择来选择了不同个数的最优特征子集,并使用 KNN 分类器对特征子集的性能进行评估。可以看出,随着特征个数的增加,KNN 分类器的准确率先逐渐增加,在特征个数为 6 时,准确率达到最高值,之后随着个数的增加,准确率不断地下降,说明了特征选择方法不仅仅能够降低特征向量的维度,还能提高分类的性能。基于支持向量机的递归删除

算法(SVM-RFE)选择出的特征子集分类性能曲线明显地高于另外两个算法。所以,这三种特征选择算法中,SVM-RFE 选出的特征子集的分类性能要优于另外两个特征选择算法。

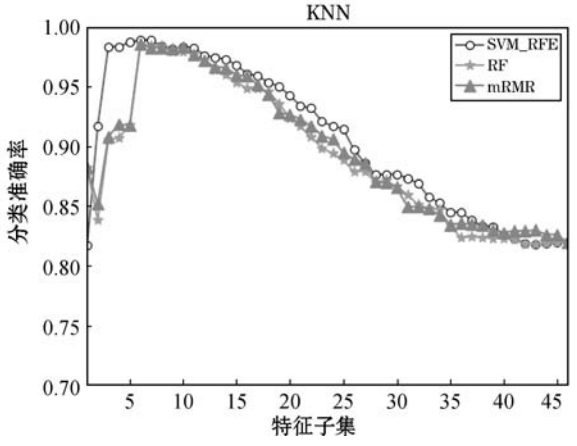

图 1 三种特征选择算法选择的特征子集的分类准确率

如图 2 所示,随着特征个数的不断增加,分类运算时间也在不断的增加。在特征个数最多的时候,运算时间也达到最高值。由此可见,特征选择算法大大提高了分类效率。针对全球恐怖主义数据库数据量大的特点,这将十分有利于对数据进行分类预测或者更复杂的分析处理。

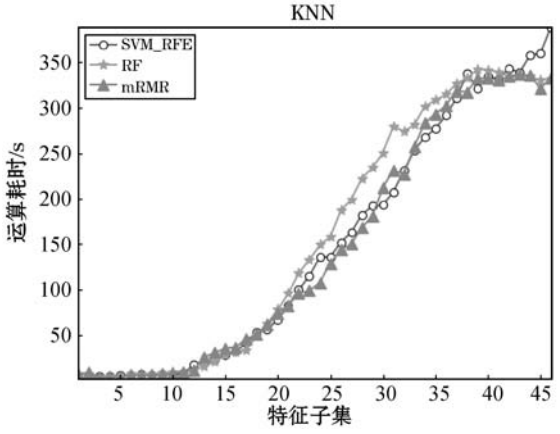

图 2 三种特征选择算法选择的特征子集的分类运算时间

对比三种特征选择算法的分类准确率的最大值和平均值,如表 1 所示,基于支持向量机的递归删除算法(SVM-RFE)均优于最小冗余最大相关算法(mRMR)和基于随机森林的特征选择算法。

表 1 特征选择算法性能比较

方法	分类准确率最大值/%	分类准确率平均值/%
SVM-RFE	98.92%	90.60%
RF	98.59%	89.21%
mRMR	98.59%	89.61%

4 结 语

本文基于全球恐怖主义数据库,针对恐怖主义组织预测研究了特征选择方法,采用了三种典型的特征选择方法。实验结果表明,基于支持向量机的递归删除算法(SVM-RFE)针对全球恐怖主义数据库的高维性、冗余性和数据不完整性的特点,经过其特征选择的特征子集能够表现出更好的分类性能。通过对比发现,特征选择方法,不仅仅能够降低特征向量的维度,还能提高分类性能和分类器的学习效率。这将有利于快速高效地处理全球恐怖主义数据库中的数据,同时也能够提高反恐怖主义活动预测的性能和效率。关于全球反恐实践技术的研究才刚刚起步,需要更多地人一起深入探索。人工智能和恐怖主义数据的相结合将有助于反恐课题的研究,未来的工作将更多的放在这个方向。

参 考 文 献

[1] 边肇祺, 张学工. 模式识别(第二版)[M]. 北京:清华大学出版社, 2000.

[2] 周松青. 全球恐怖主义数据库及对中国反恐数据库建设的启示[J]. 情报杂志, 2016,35(9):6-11.

[3] 莫豪文. 数据挖掘方法在反恐预警中的应用[D]. 北京:北京工业大学, 2017.

[4] 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). IEEE, 2017:239-244.

[5] Gundabathula V T, Vaidhehi V. An efficient modelling of terrorist groups in India using machine learning algorithms [J]. Indian Journal of Science & Technology, 2018, 11(15):1-10.

[6] Iqbal R, Murad M A A, Mustapha A, et al. An experimental study of classification algorithms for crime prediction[J]. Indian Journal of Science & Technology, 2013, 6(3):4219-4225.

[7] 李梅, 李亦农, 王玉. 信息论基础教程[M]. 北京:北京邮电大学出版社, 2015.

[8] 张睿, 马建文. 一种 SVM-RFE 高光谱数据特征选择算法[J]. 武汉大学学报:信息科学版, 2009, 34(7):834-837.

[9] 周志华. 机器学习:Machine learning[M]. 北京:清华大学出版社, 2016.

[10] Breiman L. Random forests, machine learning 45[J]. Journal of Clinical Microbiology, 2001, 2:199-228.

[11] 刘家锋, 赵巍, 朱海龙. 模式识别:Pattern recognition [M]. 哈尔滨:哈尔滨工业大学出版社, 2014.