基于随机森林的犯罪预测模型

卢　睿[1]　李林瑛[2]

（1 辽宁警察学院信息系　辽宁　大连　116036；2 大连外国语大学软件学院　辽宁　大连　116044）

摘　要：为深入挖掘犯罪嫌疑人的特征规律从而达到预防犯罪的目的，提出一种基于随机森林的犯罪预测模型。首先根据历史犯罪数据计算得出属性重要度的排序，并依此进行属性约简，然后利用所得的属性集合进行随机森林模型的训练从而得到犯罪预测模型。将脱敏后的犯罪数据应用于此模型，并以查准率和查全率对模型进行评价。实验结果表明，在犯罪信息噪声多、属性复杂的犯罪数据集中，该模型在准确度上优于支持向量机和朴素贝叶斯模型分类方法。

关键词：集成学习算法　随机森林　属性约简　犯罪预测模型

中图分类号：G353　　**文献标识码：**A　　**文章编号：**2095-7939（2019）03-0108-05

DOI：10.14060/j.issn.2095-7939.2019.03.015

1　引言

当前，我国的犯罪事件呈增长趋势且不断复杂，在犯罪数据上表现为数据量呈指数增长、数据形式复杂多样。而警方对犯罪大数据的应用仍处于一般性的定性和宏观分析上，缺乏实务性的定量的犯罪分析和预测应用，因此预测精度不足、实用价值较低。同时犯罪数据的不公开导致犯罪数据不易获得，也限制了犯罪预测研究的发展。与此相对的是，数据挖掘方法已经在不同领域的预测应用中表现出良好的性能。

研究表明，将犯罪案件、受害者和犯罪嫌疑人数据应用于数据挖掘，有助于发现隐藏的模式，从而为执法和决策者提供决策支持[1]。经公安部门研究发现，犯罪分子实施犯罪在很大程度上取决于某个人的一些基本属性，这些属性对在案后发现犯罪嫌疑人具有重大意义。随着以随机森林为代表的集成学习算法的性能得到普遍认同，很多研究者以随机森林方法为基础，将犯罪数据的诸多因素联系起来进行犯罪预测。文献[2]分别使用不同分类方法来预测谋杀案件数据中受害人与罪犯之间的关系，其研究结果认为通过随机森林和支持向量机方法建立二元分类问题可以获得良好的分类准确性，并且执行属性选择和使用透明决策树模型可以获得较好的树模型。文献[3]针对犯罪嫌疑人识别问题提出基于Probit模型的判定技术，采用聚类分离算法、关联算法及Probit模型的显著性水平参数发现重要属性并据此进行训练，从而得到嫌疑人风险判定模型。针对嫌疑人特征预测，文献[4]根据历史数据进行特征选择，训练基于SVM的特征预测模型，并与备选嫌疑人库进行特征相似度计算，进而预测犯罪嫌疑人。文献[5]针对刑事案件罪犯特征，提出改进的随机森林分类器。文献[6]采用随机森林算法进行犯罪信息指标集合的选择和犯罪风险预测。文献[7]使用随机森林回归来预测犯罪，并量化城市指标在凶杀案中的影响，进而通过掌握城市指标相对犯罪的重

收稿日期：2018-11-23

基金项目：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项目（编号：JXJZXTCX-029）；辽宁省科学研究青年项目（编号：LQ201787002）；辽宁省科学研究一般项目（编号：2016jyt-lj02）。

作者简介：卢睿（1978-），女，辽宁大连人，辽宁警察学院信息系副教授，博士，主要从事数据挖掘与情报分析研究。

要性等级达到指导控制犯罪公共政策的目的。文献[8]将Benford定律与逻辑回归、决策树、神经网络和随机森林算法结合起来，在真实的西班牙法庭案件中学习洗钱罪犯的模式。文献[9]针对保险诈骗的检测问题，提出基于随机森林、主成分分析和潜在最近邻方法的多分类系统，将随机森林作为K潜在最近邻的自适应学习机制，并以基于潜在最近邻的投票机制取代多数投票机制，从而改进基分类器的差异。

本文提出了一种基于随机森林的犯罪预测模型，能够对具体涉案人员进行犯罪风险的判定与犯罪嫌疑人识别。对犯罪嫌疑人的基本属性与犯罪倾向之间的关联性进行研究，筛选出重要的特征属性；利用所选择的特征属性进行随机森林模型的训练，最终得到犯罪预测模型。针对犯罪信息噪声多、属性复杂的特点，随机森林模型在犯罪风险预测中的应用较之支持向量机和朴素贝叶斯模型表现出更好的准确性。

2　基本理论

随机森林（Random Forest，RF）是典型的集成学习方法，在以决策树为基学习器构建 Bagging集成的基础上进一步在决策树的训练过程中引入随机属性选择[10]，并根据投票机制产生最后的分类结果。RF方法对于噪声数据和存在缺失值的数据具有很好的鲁棒性和较快的学习速度，其变量重要度度量可以作为数据的属性约简方法，所以近年已经被广泛应用到各类分类、回归、预测、特征选择及异常点检测问题中[11-15]。

定义1 给定组合分类模型$\{h_1(X),h_2(X),\cdots h_k(X)\}$，其间隔函数（Margin Function）定义为

$$mg(X,Y)=av_k I(h_k(X)=Y)-\max_{j\neq k} av_k I(h_k(X)=j) \quad (1)$$

其中(X,Y)表示服从随机分布的数据集，$I(\cdot)$为示性函数。间隔函数可以衡量平均正确分类数超过平均错误分类数的间隔程度。间隔值越多，说明分类预测的性能越好。

定义2 组合分类模型的泛化误差定义为

$$PE^*=P_{X,Y}(mg(X,Y)<0) \quad (2)$$

其中下标X,Y表示概率P覆盖X、Y空间。

定义3 如果森林中分类数目增加，根据大数定律，组合分类模型的泛化误差几乎处处收敛于

$$P_{X,Y}\{P_\theta(h(X,\theta)=Y)-\max_{j\neq Y} P_\theta(h(X,\theta)=j)<0\} \quad (3)$$

其中θ表示单棵决策树的参数向量。式（3）说明随机森林不会随着决策树的增加而产生过拟合问题，并且对于噪声和离散点是鲁棒的，缺点是可能产生一定范围内的泛化误差。

通过在袋外数据（Out of Bag, OOB）中对属性值进行扰动可以判断属性对分类结果的影响，影响越大，则说明该属性越重要。

定义4 属性j在第m（$m=1,2,\cdots,M$）棵数的属性重要度（Attribute Importance, AI）计算公式为

$$AI_m(j)=\frac{\sum_{i\in L_m} I(y^i=y_i^m)}{|L_m|}-\frac{\sum_{i\in L_{m,j}} I(y^i=y_{i,j}^m)}{|L_{m,j}|} \quad (4)$$

其中L_m表示第m棵树的袋外数据，y_i^m表示树m中属性j被扰动前的预测结果；$L_{m,j}$表示属性j被扰动后形成的新的袋外数据，$y_{i,j}^m$表示树m在$L_{m,j}$的预测分类结果；y^i为袋外数据L_m中第i个样本的实际分类值。若属性j未出现在树m中，则认为$AI_m(j)=0$。

定义5 基于OOB分类准确率的属性重要度度量，定义为OOB自变量发生轻微扰动后的分类正确率与扰动前平均分类正确率的平均减少量（Mean Decrease Accuracy, MDA），MDA计算公式为

$$MDA=\frac{1}{M}\sum_{m=1}^{M} AI_m(j) \quad (5)$$

公式（5）说明属性重要度对分类模型的贡献，以该定义作为属性约简的启发信息。

3　基于属性约简的犯罪嫌疑人分类方法

犯罪嫌疑人特征是犯罪案件特征的一部分，其分析过程需与犯罪案件特征相关联。本文构造案件基本特征与犯罪嫌疑人犯罪倾向的判定模型，分为属性约简、判定模型训练和嫌疑人犯罪倾向预测3个部分。

3.1　模型判定原理

在数据集进入方法运算之前需要做预处理，使训练集和测试集中的各个属性具有统一的定义和标准，即将与预测操作无关的冗余数据属性去除，同时也对属性值进行泛化操作、处理缺失值等，目的是提高数据质量使之适合模型的输入和运算需求。

属性约简是预测方法中的重要步骤，通过计算属性重要度将与预测结果关联较小的属性去除，只保留其中的重要属性参与运算，从而减小算法计算量、提高算法实用性。

训练数据属性约简后进入模型训练过程。本文设计了基于随机森林的训练方法，从而得到犯罪嫌疑人判定模型。

在犯罪嫌疑人预测阶段，将经过预处理后的测试数据输入预测模型，计算得出每个测试集样本的犯罪倾向，从而得出判定结论。模型的判断方法和过程如图1所示。

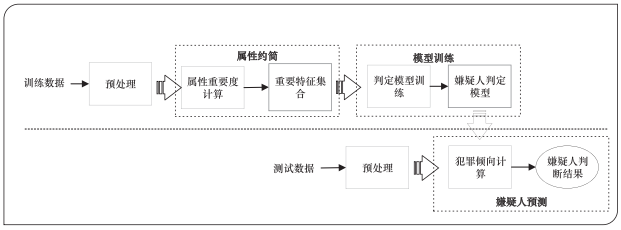

图1　犯罪嫌疑人分类方法

3.2　基于随机森林的预测模型

图2描述基于随机森林的预测模型，其中属性约简阶段采取以下步骤：

（1）以bootstrap方式生成M个自助样本集，每次未被抽取的样本构成M袋外数据。

（2）利用每个样本集m（$m=1,2,...,M$）：构造形成M个决策树T_m，其对应的袋外数据集记为L_m。

（3）运用决策树T_m对数据集L_m进行分类并记录分类结果。

（4）逐个提取每个袋外数据集实施属性值的扰动：对于每个属性$j, j=1,2,...,J$，扰动袋外数据集L_m中的属性j的取值，从而形成扰动后的数据集$L_{m,j}$。

（5）将决策树T_m应用于$L_{m,j}$进行分类运算并记录分类结果$y_{i,j}^m$。

（6）当完成对每个袋外数据集的属性值扰动后，利用公式(4)和公式(5)计算每个属性j的属性重要度。

（7）依各属性的重要度进行降序排列。

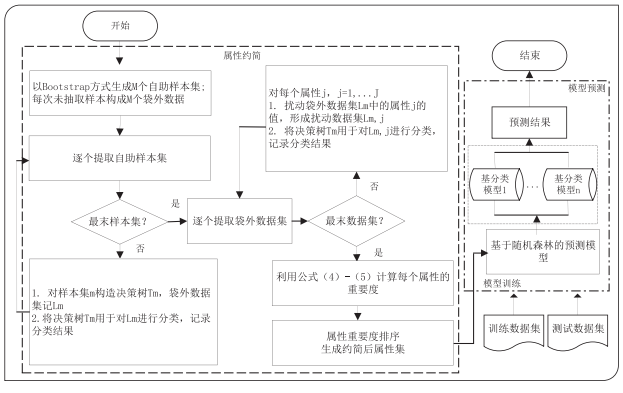

图2　基于随机森林的预测模型

对排序结果采用序列后向搜索策略进行属性约简，即每次遍历仅删除一个重要性最低的属性，产生新的特征属性集合，经过多次迭代选出最小冗余、性能最优的重要属性集合，并将其输入预测模型。

在模型训练和模型预测阶段，以随机森林思想和方法构建预测模型。在训练阶段，训练数据集进入模型进行属性约简，然后应用随机森林方法进行模型训练，从而产生n个基分类模型。将测试数据集输入各个基分类模型进行分类，然后以投票的方式决定产生预测结果。

4　随机森林实验

本文的实验数据来源于已经脱敏的犯罪人员信息的部分记录，用于挖掘犯罪嫌疑人属性特征与犯罪风险之间的证据关系，从而获得高可疑度的犯罪嫌疑人，最终达到犯罪预防和辅助决策的目的。

模型的输入信息为犯罪人员信息特征，包括年龄、家庭情况、文化程度、有无职业、有无犯罪纪录、有无特长、是否常驻人口、性别、身高、体重、经济状况。其中文化程度细分为小学、初中、高中、学士、硕士、博士等类别。模型的输出信息是对犯罪嫌疑人“犯罪程度”的分类结果，即分为{一般，严重}两类。

本文实验环境：①软件条件：MyEclipse 8.5，Weka 3.6。②硬件条件：Intel（R）Core（TM）i7-5500U @ 2.40GHz, 8GB内存，1TB硬盘，Window 7操作系统。

4.1　数据预处理

数据预处理是提高数据质量的关键步骤之一。根据实验数据的特点，需要处理数据集中的缺失值，原则上尽可能地填充缺失值，对无法填充缺失值的记录作删除处理。以“年龄”属性为例，其缺失值可通过“案发时间”和“出生日期”的差值填充。对包含多个无序不同属性值的属性向上泛化，如将“年龄”属性的特征值量化，以分组的方式划分为3个区段：{18-29}为少年，{30-40}为青年，{40以上}为中老年，相应的特征值为1~3。对于数据属性中与预测结果无关的冗余属性，如“案件ID”等，需将其删除以提高属性约简和分类运算的效率。对于各属性值中量纲和单位的不同，需要将样本数据作归一化处理，去除其对分类运算结果的影响，使处理后的数据在[0,1]区间。经过数据预处理，最终提取有效记录2021条，其中“一般”类别

1036条，“严重”类别985条，量化后的部分数据如表1所示。

表 1　犯罪人员属性值的部分量化结果

年龄	家庭情况	文化程度	有无职业	有无犯罪纪录	有无特长	是否常驻人口	犯罪程度
3	1	1	0	1	1	0	1
2	2	2	1	0	1	1	1
2	3	4	1	1	0	1	1
1	3	2	1	1	1	1	0
3	2	1	1	1	1	1	1

4.2　属性约简

利用3.2所述方法对样本数据进行属性约简，得到各属性的MDA值。表2给出经过计算得到的12个属性{A1，A2，A3，A4，A5，A6，A7，A8，A9，A10，A11，A12}的MDA值。经过计算和约简得出{A1，A2，A3，A6，A8，A9，A10}为重要属性。为便于比较，图3给出将约简的重要属性值分别除以其最大值后的结果。

表 2　属性重要性度量

属性编号	属性	MDA
A1	年龄	0.04612
A2	家庭情况	0.02573
A3	文化程度	0.03971
A4	性别	0.00723
A5	身高	0.00111
A6	有无职业	0.03489
A7	体重	0.00236
A8	有无犯罪纪录	0.06211
A9	有无特长	0.02001
A10	是否常驻人口	0.01858
A11	民族	0.00187
A12	经济状况	0.00512

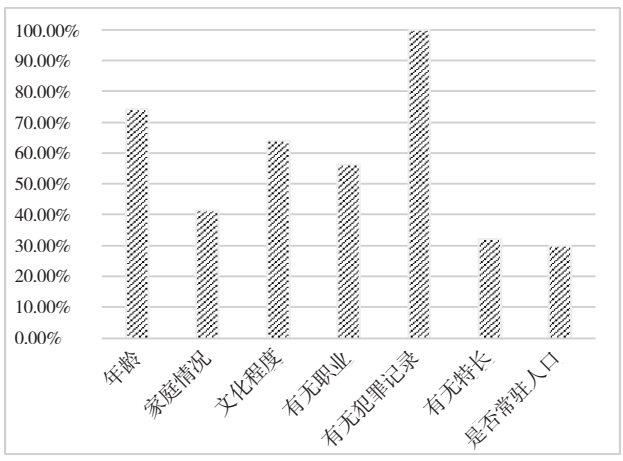

图 3　属性特征约简结果

4.3　结果与分析

根据最终确定的重要属性，约简原数据中冗余的属性列，余下的数据构建预测模型的数据集，并采用10-折交叉验证。采用控制变量法调参以使预测获得较好准确率，参数优化结果见表3，可知参数最终确定为：森林中树的棵数设为200，每次分裂随机选择的候选变量个数为3。

表 3　随机森林模型参数设置及相应结果

变量数	棵数	准确率
1	200	0.8691
2	200	0.8732
3	200	0.8803
4	200	0.8627
5	200	0.8602
2	50	0.8636
2	100	0.8711
2	200	0.8736
2	300	0.8661
2	400	0.8686

模型的查准率P和查全率R可以作为衡量模型性能优劣的指标。综合考虑查准率和查全率，可以使用F1度量，其含义是加权调和平均值。现实应用中要求漏查嫌犯的数量尽量小，因此查全率更为重要。令TP、FP、TN、FN分别表示真正例、假正例、真反例、假反例的样例数。F1度量的一般形式为F_β，能够表达出对查准率和查全率的不同偏好，其计算公式为

$$P=\frac{TP}{TP+FP} \tag{6}$$

$$R=\frac{TP}{TP+FN} \tag{7}$$

$$F_\beta=\frac{(1+\beta^2)\times P\times R}{(\beta^2\times P)+R} \tag{8}$$

其中$\beta>0$度量查全率对查准率的相对重要性，$\beta>1$时对查全率有更大影响，本文设置β=1.5。

此次实验的最终结果如表4所示。

为验证随机森林预测模型的性能，在Weka平台上分别选用SVM单分类器算法和朴素贝叶斯单分类器算法，并以默认参数进行运算，结果的比较如图4所示。可见随着输入特征变量的增多，三类算法的查准率逐渐提高，说明在一定范围内，模型的输入变量越多，预测效果越好。随机森林算法的查准率明显优于SVM单分类器算法和朴素贝叶斯单分

类器算法。原因是集成学习算法能够通过综合不同基分类器模型的分类结果来增强集成学习算法的容错性和泛化能力。表4和图4的数据说明了所提出的嫌疑人预测模型的可行性，通过该模型可以预测新发生案件中的高危犯罪嫌疑人，分析结果可进一步在相关数据库中碰撞比对，从而实现重点研判、提高办案效率的目的。

表 4　随机森林模型预测结果

实验次数	查准率	查全率	F_β
1	0.8654	0.7872	0.8097
2	0.8611	0.8123	0.8267
3	0.8824	0.8762	0.8781
4	0.8378	0.7817	0.7981
5	0.8735	0.8201	0.8358
6	0.8619	0.8163	0.8298
7	0.8731	0.7873	0.8118
8	0.8661	0.88521	0.8792
9	0.8772	0.87152	0.8733
10	0.8315	0.7809	0.7958

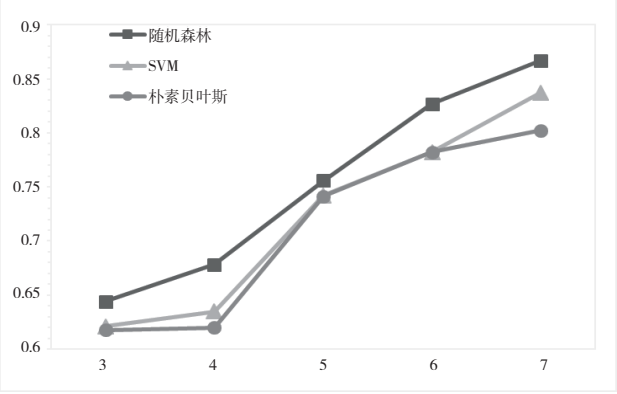

图 4　不同模型的预测效果比较

5　结论

对犯罪嫌疑人进行有效预测，不仅实现快速打击，还达到犯罪预防的目的。集成学习算法已经在不同邻域的预测应用中表现突出。本文提出基于随机森林的犯罪嫌疑人预测模型，对犯罪嫌疑人的属性加以评价和约简，有效提高了方法效率和准确性，避免了单一决策树分类的局限性。通过脱敏案件数据对模型进行评价，结果显示所提出的模型较SVM和朴素贝叶斯方法具有更好的准确性，模型可进一步应用于不同类别案件的犯罪嫌疑人预测应用中。

参考文献：

[1]唐德权,史伟奇,凌志刚.有组织犯罪集团挖掘算法研究[J].中国刑警学院学报,2015(1):26–28.

[2]Yang R, Olafsson S. Classification for predicting offender affiliation with murder victims[J]. Expert Systems with Applications, 2011(11):13518–13526.

[3]罗森林, 刘峥, 郭亮,等. 基于Probit的犯罪嫌疑人判定方法研究[J]. 北京理工大学学报, 2011(11):1337–1341.

[4]李荣岗,孙春华,姬建睿.基于支持向量机的嫌疑人特征预测[J].计算机工程, 2017(11):198–203.

[5]孙菲菲,曹卓,肖晓雷.基于随机森林的分类器在犯罪预测中的应用研究[J].情报杂志, 2014(10): 148–152.

[6]王雨晨,过仲阳,王媛媛.基于随机森林的犯罪风险预测模型研究[J].华东师范大学学报(自然科学版), 2017(4):89–96.

[7]Alvesa L G. A,Ribeirob H V,Rodriguesa F A.Crime prediction through urban metrics and statistical learning[J].Physica A:Statistical Mechanics and its Applications,2018(505): 435–443.

[8]Badalvalero E, Alvarezjare□o J A, Pavía J M. Combining Benford’s Law and machine learning to detect money laundering. An actual Spanish court case[J]. Forensic Science International, 2017(282):24–34.

[9]Lia Y, Yana C, Liub W ,et al. A principle component analysis–based random forest with the potential nearest neighbor method for automobile insurance fraud identification [J]. Applied Soft Computing, 2018(70): 1000–1009.

[10]周志华.机器学习[M]. 北京:清华大学出版社,2016:179–180.

[11]姚登举,杨静,詹晓娟.基于随机森林的特征选择算法[J].吉林大学学报(工学版), 2014(1):137–141.

[12]王慧,郭红涛.基于约简决策表的网络犯罪行为关联分析[J].中国人民公安大学学报(自然科学版), 2015(2):67–70.

[13]Chen W, Xie X, Wing J, et al. A comparative study of logistic model tree, random forest, and classification and regression tree models for spatial prediction of landslide susceptibility[J]. Catena, 2017(151): 147–160.

[14]Matin S S, Farahzadi L, Makaremr S, et al. Variable selection and prediction of uniaxial compressive strength and modulus of elasticity by random forest[J]. Applied Soft Computing, 2018(70): 980–987.

[15]Hapfelmeier A,Uim K.A new variable selection approach using Random Forests[J]. Computational Statistics and Data Analysis, 2013(60) :50–69.

（责任编辑：于　萍）