基于XGBoost算法的预警模型研究

陆万万，王维芳，马煜敏
（上海计算机软件技术开发中心，上海 201112）

摘要：针对智慧警务建设过程中犯罪预警机制滞后的问题，提出了一种融合警务系统中人员数据的通用预警模型。该模型采用随机森林算法对高维稀疏样本特征进行重要性排序，筛选得到最优特征子集。利用SMOTE过采样算法对训练集样本进行采样处理，以平衡正负样本集。该文预警模型基于XGBoost算法实现风险样本数据的分类提取，并使用粒子群优化算法对XGBoost模型的AUC值为目标函数做参数调优，提高模型预测精度。结果表明，该模型在不均衡数据集下平均准确度可达到90%以上。

关键词：大数据；XGBoost；随机森林（RF）；SMOTE过采样算法；粒子群算法（PSO）
中图分类号：TN0 **文献标识码：**A **文章编号：**1674-6236（2022）19-0049-06
DOI：10.14022/j.issn1674-6236.2022.19.011

Research on early warning model based on XGBoost algorithm

LU Wanwan, WANG Weifang, MA Yumin
（*Shanghai Computer Software Technology Development Center*, *Shanghai* 201112, *China*）

Abstract: Aiming at the problem of lagging crime early warning mechanism in the construction of smart policing, a general early warning model is proposed that integrates personnel data in the police system. The model uses the random forest algorithm to rank the importance of high-dimensional sparse sample features, and screens to obtain the optimal feature subset. The SMOTE oversampling algorithm is used to sample the training set samples to balance the positive and negative sample sets. The early warning model in this paper is based on the XGBoost algorithm to realize the classification and extraction of risk sample data, and uses the particle swarm optimization algorithm to optimize the parameters of the AUC value of the XGBoost model to improve the model's prediction accuracy. The results show that the average accuracy of the model can reach more than 90% under unbalanced data sets.

Keywords: big data; XGBoost; Random Forest (RF); SMOTE oversampling algorithm; Particle Swarm Optimization (PSO)

在以前的警务工作中，犯罪线索多来源于线下举报，随着大数据技术的发展，警务信息化程度不断提高，为线上数据实时犯罪线索预警提供了可能性。犯罪预警基于对“五要素”进行研判分析，即人、事、地、物、组，从已有的历史警务数据中挖掘出潜在风险的犯罪人员，实现对犯罪活动的预知、预警、可防、可控[1-3]。

当今大数据时代下，警务数据资源丰富，数据资源采集、处理及存储技术为犯罪预警提供了技术支撑。随着大数据技术在警务工作中的深入应用，相关领域研究者在电信诈骗侦察、社会治安治理、预警恐怖犯罪活动等方面开展了大量大数据警务应用研究，以提高警务治理能力[4-8]。例如，中国人民公安大学的陈鹏等人引入机器学习识别风险人员身份特征，通过该模型以及二项逻辑回归算法实现对风险人员的预测预警[9]。山东省科学院情报研究所的魏墨济等人提出基于网络社交媒体大数据，构建社会立场主题库，通过观点挖掘技术及分类算法判断社

收稿日期：2021-07-25 **稿件编号：**202107157
作者简介：陆万万（1984—），男，浙江宁波人，硕士研究生，工程师。研究方向：数据治理与大数据应用。

会敏感话题事件危险观点持有者的倾向性，实现了一种新型网络实时犯罪防控预警机制[10]。为提高公安机关打击涉毒犯罪活动的能力，中国人民公安大学的石一婷采用Logistic回归分析提取涉毒犯罪影响因素，构建涉毒犯罪预警模型[11]。陕西警官职业学院华艳红，采用麦克劳林公式以及泰勒公式来降低传统C4.5算法信息增益率计算过程量，对传统决策树进行剪枝修改，得到了精确度较高的预警模型[12]。对此，基于以上研究以及警务工作的现实需求，以警务大数据为基础，该文研究一种基于RF-SMOTE-XGBoost的风险人员预警模型，为犯罪预警防控工作提供技术手段支持。

1 警务数据处分析与特征提取

1.1 警务数据分析

人们在生产生活过程中，会产生丰富的社会基础数据信息。因此，在智慧警务建设过程中，建立建全以人员数据为核心且覆盖到社会各层面的社会情报资源数据库，通过对数据信息进行分析和研判，建立数据模型，可实现提前预警的目标。

该文主要从人员的自然属性、社会属性、活动情况、关联事件案、财务情况、社会关系等情况，并细分维度进行人员数据信息的分析和处理。人员数据信息表如表1所示。

表1 人员数据信息表

数据维度	详细信息
人员基本信息	身份证号码；电话号码（手机、固话）；网络账号（即时通信、社交网络、网络论坛、网络邮箱、网盘、网购账号）；职业；户籍信息（亲属关系）；金融资产账号（银行、保险等）；车牌号
动态轨迹信息	民航、铁路等订票数据；网吧、旅馆登记数据；前往重点管控区域记录
犯罪记录	犯罪类型；犯罪次数；初次犯罪年龄；最近犯罪年龄
风险行为记录	风险物品购买记录；异常转账记录；与风险人员或境外异常人员通联记录

1.2 警务数据特征提取

分析上述四个维度的数据可知，人员数据具有多源异构性，且不同的数据层面存在结构化数据和非结构化数据，因此，需采用不同的数据处理方式进行警务数据特征提取。进而，将结构化与非结构化数据处理所得的特征数据进行归一化处理，形成特征文件。警务数据特征提取流程如图1所示。

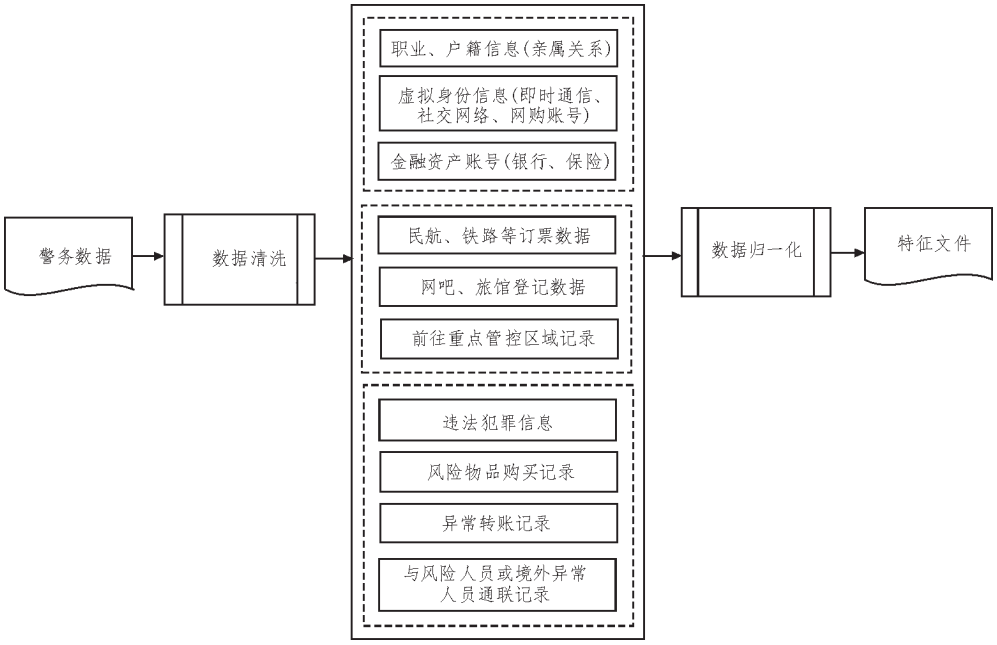

图1 警务数据特征提取流程

1）人员基本信息

人员身份证号码、电话号码（手机、固话）、网络账号（即时通信、社交网络、网络论坛、网络邮箱、网盘、网购账号）等结构化数据便于特征提取，但对于职业、地址等非结构化数据需先分类处理再进行one-hot编码。分类处理方式如下：

①所有职业归至20个职业类别中，使用相应职业类别的one-hot编码作为特征向量。

②样本的地理位置信息，例如，户籍地址、工作地点等，均采用行政区划代码来表示。

2）动态轨迹信息

该部分的数据主要包括三部分：民航、铁路等订票数据；网吧、旅馆登记数据；前往重点管控区域的记录数据。

①针对民航、铁路等订票数据：统计风险人员三年内民航、铁路订票次数及日期分布等相关特征。此外，统计风险人员前往重点管控区域的次数及日期等相关特征。

②针对住宿登记数据：统计风险人员三年内宾馆住宿或上网等相关特征。除此之外，统计前往重点管控区域住宿或上网等相关特征。

3）犯罪记录

犯罪记录主要包括该人员涉案次数等结构化数据以及犯罪描述等非结构化数据[13]。其中，犯罪描述通过归至相应犯罪类型，并对类型进行one-hot编码来量化。

4）风险行为记录

风险行为记录包括风险物品购买记录、与高危风险人员异常转账记录、与风险人员或境外异常人员通联记录三部分。

针对风险物品购买记录数据：购买风险物品危险等级采用one-hot编码来量化，统计购买风险物品次数；异常转账记录进行one-hot编码，并统计该样本与高危风险人员异常转账次数；统计样本与风险人员或境外异常人员的通联次数。

完成以上特征量化工作后，得到原始特征向量，再进行归一化处理。该文采用标准差标准化（zero-mean）数据归一化方式，处理所得数据为标准的正态分布，归一化计算公式如式（1）所示：

$$X^{*}=\frac{x-\mu}{\sigma} \tag{1}$$

式中，X^{*}为归一化所得结果值，x为样本数据，μ为样本数据均值，σ为样本数据标准差。

表2为原始样本特征数据示例。

表2　人员数据信息表

风险物品购买次数	违法数/起	异常转账次数/次	与高危人员通联次数/次	前往重点管控区域次数/次
8	3	5	1	1
2	1	3	0	0
1	0	0	0	1
4	2	4	2	0
2	1	1	0	0

风险人员原始样本归一化后，其数据无单位如表3所示。表3中，风险人员原始样本通过归一化处理后，可以尽可能消除数据不同属性对数据建模的消极影响，加快收敛速率，提高模型研判精度。

表3　风险人员原始样本归一化数据示例

风险物品购买次数/次	违法数/起	异常转账次数/次	与高危人员通联次数/次	前往重点管控区域次数/次
23.795 6	1.302 2	19.368 7	−4.995 3	0.764 8
−2.979 3	0.340 2	−3.085 0	−6.658 2	−0.573 2
−3.628 6	−0.725 7	−7.482 2	−6.658 2	0.764 8
0.230 6	0.980 4	10.582 9	−7.437 7	−0.573 2
−2.979 3	0.340 2	−5.275 5	−6.658 9	−0.573 2

2　预警模型研究

基于以上警务数据综合分析，该文提出一种基于随机森林（Random Forest，RF）、SMOTE过采样、粒子群算法（Particle Swarm Optimization，PSO）优化的极端梯度提升算法（Extreme Gradient Boosting，XGBoost），构建风险人员预警模型。

2.1　随机森林算法

预警模型建立初期需对高维稀疏的警务数据特征进行筛选过滤，因此，采用随机森林算法对警务数据特征指标的重要性进行排序，以便筛选出更具代表性的警务特征指标[14,17]。随机森林算法基于决策树理论，每次随机抽取含k个特征指标的数据子集，然后筛选其中一个最优特征指标进行划分。一般设定特征指标抽取个数k值为：

$$k=\log_2 d \tag{2}$$

其中，d为特征指标数。随机森林算法计算流程如下：

1）通过n组袋外数据测试每棵决策树性能，计算得到决策树子模型的误差值$\mathrm{error}_i(i=1,2,\cdots,n)$。

2）对n组袋外数据的第i组特征添加噪声干扰，计算得到添加噪声干扰后每棵树的误差值$\mathrm{Error}_i(i=1,2,...,n)$。

3）由以上步骤，计算可得前后两次添加噪声干扰后的误差变化平均值。

4）由于特征指标重要性与计算所得误差变化平均值呈正相关。因此，可得特征重要性公式为：

$$\mathrm{Im}_j=\frac{1}{n}\sum_{i=1}^{n}\left(\mathrm{Error}_{ji}-\mathrm{error}_{ji}\right) \tag{3}$$

5）基于特征指标重要性数值对特征指标进行重要性排序并筛选出其中重要特征指标。

2.2 SMOTE过采样算法

以上通过随机森林算法筛选得到警务特征数据集，仍存在正负样本比例失衡问题，易导致预警模型过拟合，模型泛化能力差。因此，从数据的采样方法入手，将以上筛选得到的警务特征数据集通过SMOTE过采样方法，生成少量样本来控制正负样本的数量以实现样本平衡[15-17]。

少量样本集定义为 $x_i(i=1,2,\cdots,n)$，x_i 的第 j 个属性定义为 $x_{ij}(j=1,2,\cdots,m)$。同理，若负样本集为 $y_i(i=1,2,\cdots,N)$，则 y_i 的第 j 个属性定义为 $y_{ij}(j=1,2,\cdots,m)$。

x_i 的同类K-近邻定义为 $\mathrm{NE\_}P_i=\{\mathrm{ne\_}p_{ik}|k=1,2,\cdots,K\}$，异类K-近邻样本集为 $\mathrm{NE\_}N_i=\{\mathrm{ne\_}n_{ik}|k=1,2,\cdots,K\}$，近邻候选集合为 $\mathrm{CAND}_i=\{\mathrm{cand}_{ik}|k=1,2,\cdots,K\}$。此外，当少量样本满足属于近邻候选集的条件下，允许生成新样本，同类K-近邻样本 $\mathrm{ne\_}p_{ik}$ 与 x_i 的距离定义为 $d(i,k)$。

令少量样本 $x_1\in\mathrm{NE\_}P_i$，则生成定义新的少量样本为 $e_1=[e_{11},e_{12},\cdots,e_{1m}]$，则第 j 个属性计算公式如下：

$$e_{1j}=x_{1j}+(x_{2j}-x_{1j})\cdot\mathrm{rand}[0,1] \tag{4}$$

直至达到过采样率，重复计算式(4)，得到 m 个样本子集，合成新的少量样本 e_1。

2.3 XGBoost算法

该研究预警模型基于XGBoost算法实现风险样本数据的分类提取。对经随机森林算法和SMOTE过采样算法优化后的样本数据集采用基于XGBoost的预警模型进行训练。该文集成树模型XGBoost算法通过目标函数项中含有的正则化项可有效避免算法过拟合问题[18-19]。此外，XGBoost算法对具有稀疏特征的警务数据处理效果良好，通过残差拟合多次计算得到预警结果，提高分类精度。

该文建立的XGBoost算法将各类样本集定义如下：

设样本数据集为：

$$D=\{(x_i,y_i)\}(i=1,2,\cdots n) \tag{5}$$

式(5)中，x_i 为第 i 个样本的属性集，y_i 为第 i 个样本所属类别。因此，可得第 l 棵树的预测结果为：

$$\tilde{y}_l=\sum_{k=1}^{K}f_k(x_i) \tag{6}$$

式(6)中，$f_k(x_i)$ 为第 k 棵树的预测结果，得到：

$$\mathrm{loss}=\sum_i l(\tilde{y}_l,y_i)+\sum_k\Omega(f_k) \tag{7}$$

$$\Omega(f_k)=T+\frac{1}{2}\lambda\|w\|^2 \tag{8}$$

式(7)-(8)中，$\tilde{y}_l$ 为模型的预测值；y_i 为该样本的实际值；k 为树的数量；f_k 为第 k 棵树模型；T 为该棵树叶子节点的数量；w 为在每棵树叶节点的分数；λ 为超参数。XGBoost模型的训练过程如下：

$$\tilde{y}_l^{(0)}=0 \tag{9}$$

$$\tilde{y}_l^{(1)}=f_1(x_i)=\tilde{y}_l^{(0)}+f_1(x_i) \tag{10}$$

$$\tilde{y}_l^{(t)}=\sum_{k=1}^{t}f_k(x_i)=\tilde{y}_l^{(t-1)}+f_t(x_i) \tag{11}$$

式(9)-(10)中，$\tilde{y}_l^{(t)}$ 为第 t 轮的模型预测值，其保留 $t-1$ 轮的模型预测并且在此基础上加入了一个新的函数。因此，第 t 轮的目标函数为：

$$\mathrm{loss}^{(t)}=\sum_{i=1}^{n}l(y_i,\tilde{y}_i^{(t-1)}+f_t(x_i))+\Omega(f_t) \tag{12}$$

通过泰勒展开式，取其前三项并且移除最小项，可将目标函数转化为：

$$\begin{aligned}&\mathrm{loss}^{(t)}=\sum_{i=1}^{n}[a+b+c]+d\\&a=l\left(y_i,\tilde{y}_i^{(t-1)}\right)\\&b=g_if_t(x_i)\\&c=\frac{1}{2}h_if_t^2(x_i)\\&d=\sum_{k=1}^{K}\Omega(f_t)\\&g_i=\partial_{\tilde{y}_i^{(t-1)}}l\left(y_i,\tilde{y}_i^{(t-1)}\right)\\&h_i=\partial^2_{\tilde{y}_i^{(t-1)}}l\left(y_i,\tilde{y}_i^{(t-1)}\right)\end{aligned} \tag{13}$$

此时，将叶节点的最优值代入式(13)，计算得到目标函数如下：

$$\begin{aligned}&\mathrm{loss}^{(t)}(q)=\sum_{j=1}^{T}[a+b]+\gamma T\\&a=\left(\sum_{i\in I_j}g_i\right)w_j\\&b=\frac{1}{2}\left(\sum_{i\in I_j}h_i+\lambda\right)w_j^2\end{aligned} \tag{14}$$

2.4 PSO参数优化

模型训练过程中，采用粒子群算法(Particle Swarm Optimization，PSO)对基于XGBoost算法的预警模型进行参数优化，提高模型分类精度。粒子具有速度和位置两类属性，粒子在空间中运动的快慢用速度来表示，位置的三维向量为XGBoost模型的三个超参数，即学习率、树深度以及最小叶节点权重。选定基于XGBoost算法的预警模型的AUC值作为优化目标，通过每个粒子单独追随当前搜索到的最优值来寻找全局最优。

其算法流程如下：

1)在初始化范围内，对粒子群进行随机初始化，

包括随机位置和速度。

2）根据适应度函数（Fitness Function），计算每个粒子的适应值。

3）对每个粒子，将其当前适应值与其个体历史最佳位置对应的适应值作比较，如果当前的适应值更高，则用当前位置更新粒子个体的历史最优位置 $Pbest_i$；对每个粒子，将其当前适应值与历史最佳位置对应的适应值作比较，如果当前的适应值更高，则用当前位置更新粒子群的全局最优位置 $Gbest_i$；算法的迭代公式如下：

$$\begin{aligned} x_i &= (x_{i1}, x_{i2}, x_{i3}) \\ x_i &= x_i + v_i \\ v_i &= w \cdot v_i + a + b \\ a &= c_1 \cdot \mathrm{rand}_1() \cdot (Pbest_i - x_i) \\ b &= c_2 \cdot \mathrm{rand}_2() \cdot (Gbest_i - x_i) \end{aligned} \tag{15}$$

式（15）中，x_i 为该粒子当前的位置；v_i 为第 i 个粒子速度；c_1、c_2 为学习因子；w 为惯性权重，用于平衡搜索速度和搜索精度。w 值较大，全局寻优能力强，局部寻优能力弱；反之同理。

4）若未达到终止条件，则转步骤2）。

粒子通过上述步骤对位置和速度实现迭代，逐步搜索得到最优点。

2.5　预警模型实现流程

该文所建立的预警模型算法流程如下：

1）首先，将样本集归一化处理。此外，考虑到警务数据高维稀疏特性，先去除其中缺失数据以及易导致过拟合的特征。然后通过随机森林算法对预警特征值重要性排序，并筛选得到特征数据集 S'。

2）设置训练集、验证集、测试集比例为8∶1∶1。基于数据集正负样本比例通过SMOTE过采样算法平衡样本集，生成训练集和验证集的新样本集合 S'。

3）将以上新样本集合 S' 划分为训练集 S'_{train} 与验证集 S'_{test}，在 S'_{train} 中随机抽取，得到 S'_{train1}，S'_{train2}，…，S'_{trainm}。

4）随机产生 N 组解，每组解为XGBoost的三个超参数，即学习率、树的最大深度以及最小叶子节点样本权重。

5）XGBoost模型的AUC值作为本预警模型的适应度函数 f，通过粒子群算法优化，得到最小误差 $f_{\min}$ 以及相应的最优解。

6）将测试集数据代入训练完成的预警模型加以检验，并对比其他分类算法的预测精度。

3　模型测试与分析

该文预警模型验证实验中所使用到的20 000个人员信息样本均来源于公安系统，且针对其中的业务敏感信息进行了脱敏处理，以保证数据安全性。

3.1　模型优化与结果分析

为验证该文所建立预警模型性能，分别以该文预警模型与模型未经过特征提取、模型未经过SMOTE处理、模型未经过PSO粒子群算法做参数调优三种状态进行性能对比。通过ROC曲线以及AUC值评价该文模型算法性能，AUC值为ROC曲线所覆盖的区域面积，AUC越大，分类器分类效果越好，即模型预警效果越好。上述模型ROC曲线图如图2-4所示。（TPR：在实际中为阳性的样本被判断为阳性的比例；FPR：在实际中是阴性的样本，但是判断为阳性的比例，该曲线如果是一条45°斜线时，证明模型拟合得特别准确）

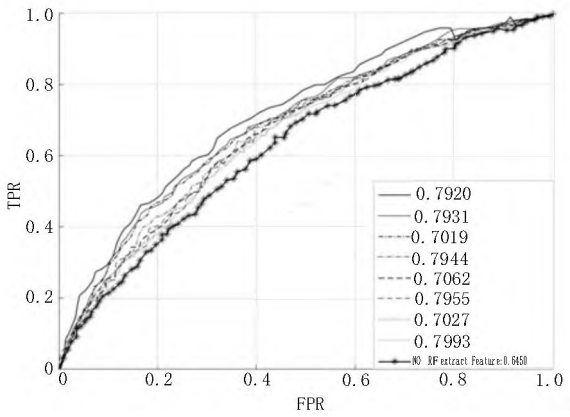

图2　未经特征提取的预警模型ROC曲线

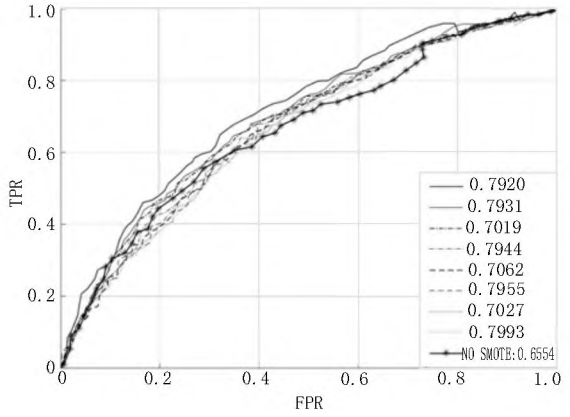

图3　未经SMOTE处理的预警模型ROC曲线

从图2-4可知，该文预警模型，即经过随机森林算法、SMOTE过采样算法、粒子群算法优化的

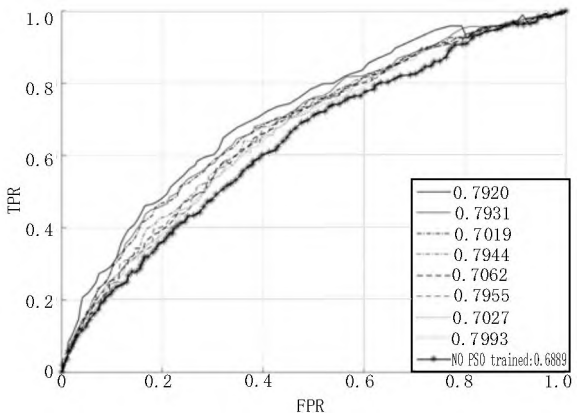

图4　未经过粒子群算法参数调优的预警模型ROC曲线

XGBoost模型AUC值为0.792 0。未经过随机森林算法特征提取AUC值为0.645 0，随机森林特征提取后AUC值同比提高23%。未经过SMOTE过采样算法平衡样本数据AUC值为0.655 4，经过SMOTE过采样处理后模型AUC值0.792 0同比提高21%。未经过粒子群算法优化的模型AUC值为0.688 9，经过粒子群优化后的XGBoost模型的AUC值0.792 0同比提高15%。因此，可得随机森林算法、SMOTE过采样算法、粒子群算法优化均可提高预警模型精度。

3.2　预警模型精确度

由于实际警务系统数据中正负样本比例失衡严重，为便于验证预警模型在不均衡数据集上的精确度，设置几组不同正负样本比例的对照实验，比例分别为：10∶1,5∶1,2∶1,1∶1。在每组不同正负样本比例的基础上，再设置四组测试集样本比例为10%、20%、30%和40%。在以上四组测试集样本比例下测得的结果取平均值作为最终结果。

此外，为验证该文基于XGBoost的预警模型精确度，设置了五组经典的机器学习算法进行结果对比，即梯度提升决策树GBDT(Gradient Boost Decision Tree)、支持向量机SVM(Support Vector Machine)、K-近邻分类器KNN(K-Nearest Neighbor)、高斯朴素贝叶斯分类器GNB(Gaussian Naive Bayes)、逻辑回归LR(Logistic Regression)。

各模型在不同正负样本比例下的准确度如表4。

从表中可以看到，在四组不同正负样本比例的实验中，该文基于RF-SMOTE-PSO-XGBoost算法的预警模型准确度最高。且正负样本比例越大，准确度越高。因此，该文模型在实际警务业务域中，即高度不均衡数据集下平均准确度可达到90%以上。该文基于RF-SMOTE-PSO-XGBoost算法的预警模型构建过程中使用到的随机采样过程很大程度上提高了模型的泛化能力，使得模型在不同数据比例下均保持相对稳定的表现。

表4　风险人员原始样本归一化数据示例

模型	正负样本比例			
	10∶1	5∶1	2∶1	1∶1
RF-SMOTE-PSO-XGBoost	93.85%	90.49%	85.79%	87.65%
GBDT	92.05%	83.02%	78.88%	85.30%
SVM	92.54%	87.49%	80.97%	80.04%
KNN	91.57%	86.12%	80.33%	79.83%
GNB	92.13%	86.55%	75.12%	64.07%
LR	93.08%	89.74%	84.47%	83.95%

4　结　论

该文建立了一种基于XGBoost算法的预警模型。针对警务大数据样本集，采用随机森林算法进行冗余度筛选以及SMOTE算法平衡正负样本比例。另外，通过粒子群算法优化基于XGBoost的预警模型。通过设置对照实验，该文预警模型所采用算法均提高了模型预警精度。同时，该文基于XGBoost的预警模型相较于大部分同类型算法计算速度快、准确性高且模型的泛化能力较好。因此，该文预警模型对今后公安系统潜在风险人员的数据挖掘研究具有一定的借鉴意义。

参考文献：

[1] 李忠东.预测犯罪[J].检察风云,2019(5):32-33.
[2] 孙自强,于龙.公安大数据时代金融犯罪预警防控研究[J].中国防伪报道,2020(12):82-89.
[3] 王海林,井晓龙,魏慧雯."从人到案"侦查模式在电信网络诈骗犯罪侦查中的应用[J].中国人民公安大学学报(自然科学版),2020,26(4):74-80.
[4] 徐宗本,冯芷艳,郭迅华,等.大数据驱动的管理与决策前沿课题[J].管理世界,2014(11):158-163.
[5] 邓玉洁,康洺晞.大数据在电信诈骗案件侦查中的应用[J].警学研究,2020(2):69-79.
[6] 吴跃文.大数据背景下跨境电信网络诈骗犯罪的预警与反制——以冒充公检法诈骗为例[J].湖北警官学院学报,2019(3):89-96.
[7] 袁春瑛.大数据思维视域下的社会治安治理方式创新[J].山东警察学院学报,2020,32(2):134-140.
[8] 李志恒,姚博.大数据预警恐怖活动犯罪流程及实

（下转第59页）

[J].情报学报,2019,38(3):266-276.
[5] 苏宣瑞,邹秀清,丁勇.一种基于区块链的身份识别技术[J].中兴通讯技术,2018,24(6):41-48.
[6] 张建文,程海玲."破碎的隐私承诺"之防范:匿名化处理再识别风险法律规则研究[J].西北民族大学学报(哲学社会科学版),2020(3):76-86.
[7] 张勇.个人信息去识别化的刑法应对[J].国家检察官学院学报,2018,26(4):91-109,174.
[8] Lee K,Filannino M,Uzuner Ö.An empirical test of GRUs and deep contextualized word representations on de-identification[J].Studies in Health Technology and Informatics,2019,264(5):218-222.
[9] 吴梦婷,孙丽萍,刘援军,等.基于约束聚类的k-匿名隐私保护方法[J].计算机工程与设计,2021,42(3):607-613.
[10]吴奇烜,马建峰,孙聪.采用完整性威胁树的信息流完整性度量方法[J].网络与信息安全学报,2019,5(2):50-57.
[11]王睿,陈立全,沙晶,等.基于双向认证的RFB远程安全数字取证方案[J].南京邮电大学学报(自然科学版),2017,37(3):106-112.
[12]Abdelaal M A,Ebrahim G A,Anis W R.High availability deployment of virtual network function forwarding graph in cloud computing environments [J]. IEEE Access,2011(12):1234.
[13]Ma Y,Liang W,Huang M,et al.Virtual network function service provisioning in MEC via trading off the usages between computing and communication resources[J].IEEE Transactions on Cloud Computing,2011(1):1-9.
[14]康朋飞.GPS/BDS组合系统伪距联合单点定位算法的对比分析[J].电子设计工程,2019,27(19):127-130,135.
[15]雍龙泉,贾伟,黎延海.基于光滑逼近函数的高阶牛顿法求解凸二次规划[J].科学技术与工程,2021,21(6):2151-2156.
[16]康茜,晏慧,雷建云.基于数据隐私保护的(L,K,d)算法[J].中南民族大学学报(自然科学版),2020,39(5):517-523.

(上接第54页)

现路径[J].辽宁公安司法管理干部学院学报,2019(4):19-24.
[9] 陈鹏,曾昭龙,胡啸峰,等.基于机器学习的犯罪人惯犯身份预测分析和识别[J].中国刑警学院学报,2018(5):124-128.
[10]魏墨济,赵燕清,朱世伟,等.基于立场建模的网络犯罪预警研究[J].计算机工程与科学,2021,43(1):151-160.
[11]石一婷.强制戒毒人员回归社会后犯罪预警模型研究[D].北京:中国人民公安大学,2020.
[12]华红艳.基于数据挖掘聚类分析的犯罪预警方法研究[J].信息技术,2020,44(12):38-42.
[13]高伟.大数据支持下的电子商务用户画像构建思考[J].数字化用户,2019,25(47):73.
[14]Yadav C S,Sharan A.Feature learning using random Forest and Binary logistic regression for ATDS[J]. Algorithms for Intelligent Systems,2020(5):341-352.
[15]Wong G Y ,Leung F,Ling S H.A novel evolutionary preprocessing method based on over-sampling and under-sampling for imbalanced datasets[C].Conference of the IEEE Industrial Electronics Society. IEEE,2014.
[16]Barua S ,Islam M M ,Murase K.A novel synthetic minority oversampling technique for imbalanced data set learning[C].Springer-Verlag,2011.
[17]Georgios D,Fernando B,Felix L.Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE[J].Information Sciences,2018,465:1-20.
[18]王名豪,梁雪春.基于CPSO-XGBoost的个人信用评估[J].计算机工程与设计,2019,40(7):1891-1895.
[19]Qin C,Zhang Y,Bao F,et al.XGBoost optimized by adaptive particle swarm optimization for credit scoring[J].Mathematical Problems in Engineering,2021(5):1-18.