计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.190600162

FS-CRF:基于特征切分与级联随机森林的异常点检测模型

刘振鹏[1,2] 苏 楠[1] 秦益文[3] 卢家欢[1] 李小菲[2]
1 河北大学网络空间安全与计算机学院 河北 保定 071002
2 河北大学信息技术中心 河北 保定 071002
3 兰州交通大学电子与信息工程学院 兰州 730070
(lzp@hbu.edu.cn)

摘 要 大数据时代,攻击篡改、设备故障、人为造假等原因导致海量数据中潜藏着许多异常值。准确地检测出数据中的异常点,实现数据清洗,至关重要。文中提出一种结合特征切分与多层级联随机森林的异常点检测模型(outlier detection model based on Feature Segmentation and Cascaded Random Forest,FS-CRF)。利用滑动窗口与随机森林对原始特征进行细粒度切分,生成类概率向量,用于训练多层级联的随机森林;由级联层中最后一层的随机森林投票决定样本的最终类别。仿真实验结果表明,新方法在基于多个UCI数据集进行的异常分类任务中均获得较高F1-measure评分;级联结构使新模型相比于经典的随机森林算法进一步提高了泛化能力;在高维数据集上所提方法比梯度提升决策树和XGBoost拥有更优的性能,且超参数较少,易于调优,具有更好的综合性能。

关键词:数据清洗;细粒度特征;级联随机森林;集成学习;异常点检测
中图法分类号 TP301

FS-CRF:Outlier Detection Model Based on Feature Segmentation and Cascaded Random Forest

LIU Zhen-peng[1,2],SU Nan[1],QIN Yi-wen[3],LU Jia-huan[1] and LI Xiao-fei[2]
1 School of Cyber Security and Computer,Hebei University,Baoding,Hebei 071002,China
2 Information Technology Center,Hebei University,Baoding,Hebei 071002,China
3 School of Electronic and Information Engineering,Lanzhou Jiaotong University,Lanzhou 730070,China

Abstract In the era of big data,there are many abnormal values hidden in massive data due to attack tampering,equipment failure,artificial fraud and other reasons. Accurately detect outliers in data is critical to data cleaning. Therefore,an outlier detection model combining feature segmentation and multi-level cascaded random forest (FS-CRF) is proposed. Using the sliding window and the random forest to segment the original features,the generated class probability vector is used to train the multi-level cascaded random forest. Finally,the category of the sample is determined by the vote of the last layer. Simulation experiment results show that the new method can effectively detect outlier in classification tasks on UCI data sets,with high F1-measure scores obtained on both high and low dimensional data sets. The cascade structure further improves the generalization ability of the model compared to the classical random forest. Compared with the GBDT and XGBoost,the proposed method has performance advantages on high-dimensional data sets,and has fewer hyper-parameters that easy to tune and has better comprehensive performance.

Keywords Data cleaning,Grained feature,Cascade random forest,Ensemble learning,Outlier detection

1 引言

异常点在数据挖掘以及统计分析中也被称为离群点或不一致点,异常点检测是找出行为不同于预期对象的检测点的过程。在数据安全的工作中开展异常点检测是发现与常规数据模式显著不同的数据模式,并分析其中潜在的数据异常,继而实现对数据的清洗,这对于保障数据安全具有重要意义。异常点检测在学术界与工业界都是研究的热点,如医疗系统中的疾病模式、金融领域中的信用卡反欺诈、安全领域中的网络入侵检测以及城市交通管理中的异常检测领域等[1-2]。另外,环境监测数据中的虚假数据泛滥,如何有效检测出这些虚假的异常类数据,也成为一个新的关注点。

许多国内外研究者已经相继对异常点检测技术作了研究。Dominguez[3]和Wang[4]等对异常点检测的方法论进行

到稿日期:2019-06-26 返修日期:2019-08-06 本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:河北省自然科学基金(F2019201427);教育部"云数融合科教创新"基金(2017A20004)
This work was supported by the Natural Science Foundation of Hebei Province,China (F2019201427) and Ministry of Education Fund for "Integration of Cloud Computing and Big Data, Innovation of Science and Education",China (2017A20004).
通信作者:李小菲(lixiaofei@hbu.edu.cn)

了全面的分析，将关键算法归为基于线性方法[5]、基于距离和密度的方法[6-7]以及基于神经网络的方法[8-9]等。近年来，集成学习中的 Bagging 方法与 Boosting 弱分类器增强方法成为了异常点检测研究的热点方法。Liu 等[10]提出的孤立森林算法，根据叶子节点到根节点的路径长度建立异常指数，对全局异常点检测的效果较好，但是不擅长处理局部的相对稀疏点。Friedman[11]提出的梯度提升决策树(Gradient Boosting Decision Tree，GBDT)，通过迭代的方式将较弱的学习器组合成一个较强的学习器；在逐次的迭代中使得残差不断减小，从而产生纵向加深的树。该方法具有预测精度高且对异常值强鲁棒等优点。Chen 等[12]提出的极端梯度提升树(Extreme Gradient Boosting，XGBoost)算法，也是按照损失函数的负梯度方向提升，只是把经验误差二阶泰勒展开，通过加入一些正则项，使损失函数具有可扩展性、高精确度及拟合效果好等特点；但该算法参数过多，使用效果过于依赖调参结果。

为了进一步提高基于决策树的集成学习方法在异常检测方面的性能和实用性，本文提出了一种基于特征切分与级联随机森林的异常点检测模型(FS-CRF)。该模型将堆叠级联结构引入随机森林，以细粒度提取数据的特征空间，使得它在多类异常数据的检测识别中具有更好的准确率。

2 异常点检测模型 FS-CRF

FS-CRF 异常检测模型如图 1 所示，包含特征切分、级联随机森林和结合策略 3 部分。特征切分对原始数据特征进行细粒度处理，级联随机森林实现数据中异常点的分类预测，结合策略部分采取投票法得出最终的异常点检测结果。

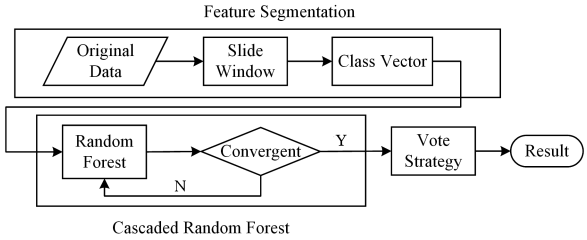

图 1 异常点检测模型

Fig. 1 Model of outlier detection

2.1 特征切分

对于无法进行大规模并行计算的单机系统来说，减小单次计算量可以避免由于进行数据异常检测而发生的系统拥塞，进而避免系统陷入异常的不安全状态。本文提出的结构对原始数据特征进行简化处理，使用滑动窗口降低单次处理的数据特征维度，减少了计算量。通过计算 OBB-袋外估计的评分(Out of Bag Estimation)，当固定窗口大小为 $n/2$(取整，n 表示原始数据的特征维度)时，切分出的特征数量既可以避免用于特征处理的单层随机森林发生过拟合，又可以降低单次处理的特征维度并生成有效类概率向量。每一个切分后的特征向量子片都将输入单层随机森林，随后生成类概率向量并进行有序重连接，形成一个重新表示的特征向量作为新的表征。特征向量切片的过程如图 2 所示。对于一个长度为 n 的一维特征向量，若使用总长度为 m 的滑动窗口进行特征切片且每次滑动一个单位长度，将产生 $n-m+1$ 个 m 维的特征向量子片。对于包含 c 个异常类别的检测问题，经过随机森林分类后，长度为 n 的一维特征向量将产生长度为 $c(n-m+1)$ 的类概率向量；同理，对于一个 $n\times n$ 的二维图像数据，将产生长度为 $c(n-m+1)^2$ 的类概率向量[13]。随后，新生成的类概率向量将作为后面级联随机森林的输入特征。图 3 阐释了一个特征向量子片在单层随机森林中生成类概率向量的过程。随机森林中的每一棵决策树都根据特征向量子片在节点上进行划分，由最后落入叶子节点上的一组值得出预测概率，然后对森林中所有决策树的预测求均值，从而得到最终输出的类概率向量。

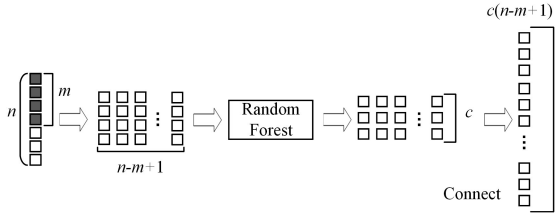

图 2 特征向量切分示意图

Fig. 2 Illustration of feature vector split

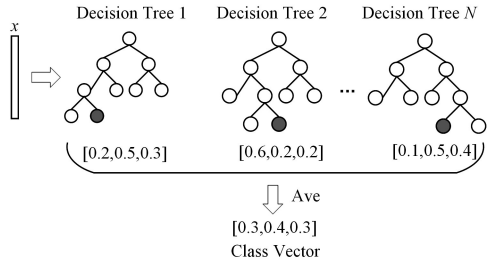

图 3 类概率向量生成示意图

Fig. 3 Illustration of class probability vector generation

2.2 级联随机森林

级联随机森林是由 CART 决策树集成的随机森林通过层级堆叠的方式形成的级联结构。级联结构中每一个新层的输入，都是由该层之前所有层的输出和原始输入聚合在一起组成的。级联随机森林的每一个级联层会统计所有 CART 决策树在输入样本上的预测结果，得出各类的比例，生成类概率向量。随后，将输入样本上预测的类概率向量与特征切分后形成的原始类概率向量拼接后作为下一个级联层的输入特征。级联随机森林相比于传统的随机森林提高了多样性，因此可以提高集成学习的泛化能力。值得说明的是，级联结构每扩展一个新的级联层，都将随机抽取 80% 的训练集作为验证集，其余 20% 作为评估集，用于验证新级联层的性能增益。当性能提升范围低于容忍度时，训练过程将终止，级联随机森林的层级数量也随之被最终确定。级联随机森林的结构如图 4 所示。

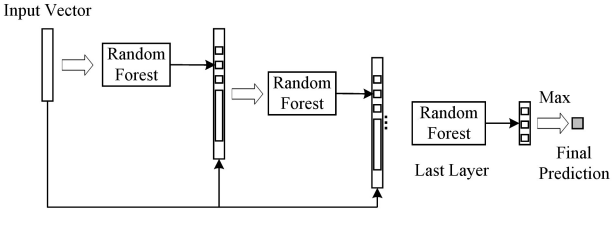

图 4 级联随机森林的结构

Fig. 4 Structure of cascaded random forest

2.3 结合策略

在集成学习中，个体学习器通过结合策略进行结合后，输出最终结果。实际的异常点检测问题可以简化为分类任务，使用投票法进行异常分类。将级联随机森林最后一层的预测结果作为整个级联随机森林的检测结果，统计最后一层森林中所有决策树的输出类别，然后在生成整个森林的类别概率分布的基础上采用投票法进行决策。学习器 h_i 从类别标记集合$\{c_1,c_2,\cdots,c_N\}$中预测出一个标记，将 h_i 在样本 x 上的预测输出表示为一个 N 维向量$(h_i^1(x);h_i^2(x);\cdots;h_i^N(x))$，其中 $h_i^j(x)$是 h_i 在类别标记 c_j 上的输出。在可靠性要求较高的异常检测任务中，采用绝对多数投票法，若某标记得票过半数则预测为该标记，否则拒绝预测。但若检测任务要求必须提供预测结果，则绝对多数投票法将退化为相对多数投票法。

绝对多数投票法：

$$H(x)=\begin{cases}c_j, & \text{if } \sum_{i=1}^{T}h_i^j(x)>0.5\sum_{k=1}^{N}\sum_{i=1}^{T}h_i^k(x)\\ \text{reject}, & \text{otherwise}\end{cases} \tag{1}$$

相对多数投票法：

$$H(x)=c_{\arg\max_j}\sum_{i=1}^{T}h_i^j(x) \tag{2}$$

2.4 算法描述

级联随机森林算法的具体步骤如算法 1 所示。

算法 1 级联随机森林算法

输入：训练集 D=$\{(x_1,y_1),(x_2,y_2),\cdots,(x_m,y_m)\}$，待测样本 x

输出：级联的随机森林集合$\{RF_i\mid i=1,2,\cdots,N\}$，待测样本 x 的投票结果 H(x)

```
Cascaded_RF={∅}
初始化参数：容忍度阈值 t 以及滑窗大小 win_size
D′=Feature Grained(D) //D′为细粒度切分得到的新数据集
do
  for i=1,2,…,T do
    对 D′做 Bootstrap，生成训练集 Di′并抽样，使用 Di′生成一棵 CART 决策树
    从 d 个特征中随机选择 k 个特征，依据 Gini 指数选择最优分裂特征，直至满足终止条件
  end for
  if(tolerance>=t)
    Cascaded_RF+={RFi}
  else
    break
while(TRUE)
H(x)=c_{argmax_j} Σ_{i=1}^{T} h_i^j(x) //级联森林最后一层投票 x 所属类别
Return Cascaded_RF
```

3 实验

3.1 实验环境及数据准备

实验平台为：Intel(R) Core(TM) i7-6500U CPU @ 2.50GHz 2.50GHz，DDR3L 8.00GB RAM；使用 python3.7 实现模型代码。选用 UCI[14] 的 4 个经典数据集对模型进行测试，数据集的相关信息如表 1 所列。对于 FS-CRF，本文使用了网格搜索方法搜寻最优参数，表 2 列出了 FS-CRF 调优参数的设置。

表 1 实验数据集的相关信息

Table 1 Information of experimental datasets

Data set	Samples	Features	Classification
Iris	150	4	3
Wine	178	13	3
Cancer	569	30	2
Digits	1797	64	10

表 2 FS-CRF 调优参数的设置

Table 2 Parameter optimized setting of FS-CRF

Parameters	Feature Grained Random Forest	Cascade Layers Random Forest
n_estimators	40	100
max_features	$\log_2 n$	$\sqrt{n}$
min_samples_leaf	4	2
min_samples_split	10%	5%
criterion	Gini	Gini

将本文算法与随机森林(RF)算法[15]、梯度提升决策树算法(GBDT)[11]以及 XGBoost 算法[12]进行对比分析。

3.2 对照实验及结果分析

表 3－表 6 列出了不同异常点检测算法在实验数据集上采用十折交叉验证法得到的对比结果，每个数据集的 80%用于训练，20%用于验证。

表 3 不同方法在 Iris 数据集上的对比结果

Table 3 Compared results of different algorithms on Iris

Algorithm	*Accuracy*	*Recall*
RF	0.96667	0.96296
GBDT	**0.98116**	0.97831
XGBoost	0.97778	**0.97875**
FS-CRF	0.97333	0.97492

表 4 不同方法在 Wine 数据集上的对比结果

Table 4 Compared results of different algorithms on Wine

Algorithm	*Accuracy*	*Recall*
RF	0.95117	0.95232
GBDT	0.95270	0.96731
XGBoost	**0.97222**	**0.97619**
FS-CRF	0.96147	0.96281

表 5 不同方法在 Cancer 数据集上的对比结果

Table 5 Compared results of different algorithms on Cancer

Algorithm	*Accuracy*	*Recall*
RF	0.94035	0.93117
GBDT	0.94736	0.91176
XGBoost	0.94281	0.93775
FS-CRF	**0.95851**	**0.94927**

表 6 不同方法在 Digits 数据集上的对比结果

Table 6 Compared results of different algorithms on Digits

Algorithm	*Accuracy*	*Recall*
RF	0.96389	0.96369
GBDT	0.96970	0.97212
XGBoost	0.97306	0.97409
FS-CRF	**0.98316**	**0.98303**

分析实验结果可以发现，在 Iris 数据集上模拟异常分类时，本文所提方法的检测效果与随机森林相当；在 Wine 数据集上，本文方法在准确率上优于 GBDT 和随机森林算法，且拥有较高的召回率；随着数据维度的提升，在 Cancer 数据集上，FS-CRF 在准确率和召回率上开始高于其他方法；在 Digits 数据集上，FS-CRF 的两项性能均已明显优于其他方法，显示出良好的异常分类性能。

使用精度和召回率的调和平均来定义 F1 评价指标，其通常被用来综合评价算法的性能。

$$F1=\frac{2\times precision\times recall}{precision+recall} \tag{3}$$

几种算法在实验数据集上的 $F1$ 结果如图 5 所示。

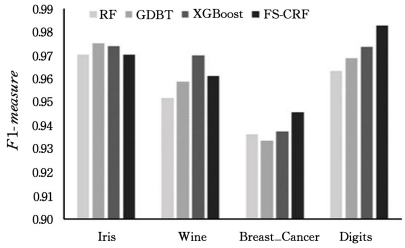

图 5 不同算法的 $F1$ 值比较

Fig. 5 $F1$ value comparison of different algorithms

从图 5 可以看出，在 Iris 数据集上，FS-CRF 方法的 $F1$ 值与 RF，GBDT 以及 XGBoost 整体相当，略低于 GBDT 和 XGBoost。在 Wine 数据集上，不同集成学习算法间出现了性能梯度差异：FS-CRF 的表现优于随机森林和梯度提升树，但仍然略低于 XGBoost。结合表 1，随着训练集特征维数的显著增加，FS-CRF 在 Cancer 数据集以及 Digits 数据集上均表现出了出色的异常类检测能力，综合评价指标优于其他 3 种方法。

综合以上数据可以看出，其他方法虽然在部分数据集上有性能优势，但在高维度数据集上的表现不如 FS-CRF 稳定；FS-CRF 在高维和低维数据集上均获得了较高的召回率，同时保证了异常分类任务的准确率，并且在高维数据集上的性能优势更加明显。

结束语 本文在传统随机森林算法的基础上提出了特征切分级联随机森林的异常点检测方法。该方法对样本特征属性进行切片处理，克服了传统方法应用于大数据时训练开销大的不足；级联结构的引入还提升了泛化能力，使算法具有较高的准确率。实验结果表明，本文方法的效果比梯度提升决策树和随机森林算法好。在实际问题中，XGBoost 超参数较多，调优复杂。本文提出的基于特征切分和级联随机森林方法的超参数少，具有更好的可实用性，在今后的异常点检测领域具有较好的研究和应用价值。

未来将会对样本特征做更多优化处理，进一步提高预测精度，并基于样本不平衡数据集的问题对算法做进一步的研究。

参 考 文 献

[1] AHMED M, MAHMOOD A N, ISLAM M R. A survey of anomaly detection techniques in financial domain[J]. Future Generation Computer Systems, 2016, 55(6): 278-288.

[2] DJENOURI Y, ZIMEK A. Outlier detection in urban traffic data [C]// Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics. ACM, 2018: 1-12.

[3] DOMINGUES R, FILIPPONE M, MICHIARDI P, et al. A comparative evaluation of outlier detection algorithms: Experiments and analyses[J]. Pattern Recognition, 2018, 74: 406-421.

[4] WANG H, BAH M J, HAMMAD M. Progress in Outlier Detection Techniques: A Survey[J]. IEEE Access, 2019, 7: 107964-108000.

[5] GUO K, LIU D, PENG Y, et al. Data-Driven Anomaly Detection Using OCSVM with Boundary Optimzation[C]// 2018 Prognostics and System Health Management Conference. IEEE, 2018: 244-248.

[6] BREUNIG M M, KRIEGEL H P, NG R T, et al. LOF: identifying density-based local outliers[J]. ACM SIGMOD Record, 2000, 29(2): 93-104.

[7] RAMASWAMY S, RASTOGI R, SHIM K. Efficient algorithms for mining outliers from large data sets[J]. ACM SIGMOD Record, 2000, 29(2): 427-438.

[8] LIU Y, LI Z, ZHOU C, et al. Generative adversarial active learning for unsupervised outlier detection[J]. arXiv: 1809. 10816.

[9] CHEN J, SATHE S, AGGARWAL C, et al. Outlier detection with autoencoder ensembles[C]// Proceedings of the 2017 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics, 2017: 90-98.

[10] LIU F T, TING K M, ZHOU Z H. Isolation-based anomaly detection[J]. ACM Transactions on Knowledge Discovery from Data (TKDD), 2012, 6(1): 1-39.

[11] FRIEDMAN J H. Greedy function approximation: a gradient boosting machine[J]. Annals of Statistics, 2001, 29(5): 1189-1232.

[12] CHEN T, GUESTRIN C. Xgboost: A scalable tree boosting system[C]// Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining. ACM, 2016: 785-794.

[13] GONG Z H, WANG J N, SU C. A Weighted Deep Forest Algorithm[J]. Computer Applications and Software, 2019, 36(2): 274-278.

[14] DUA D, GRAFF C. UCI Machine Learning Repository [EB/OL]. http://archive.ics.uci.edu/ml.

[15] BREIMAN L. Random forests [J]. Machine learning, 2001, 45(1): 5-32.

LIU Zhen-peng, born in 1966, Ph. D, professor, is a senior member of China Computer Federation. His main research inte-rests include network information security and outlier detection.

LI Xiao-fei, born in 1979, master, engineer. Her main research interests include network information security and outlier detection.