控制与决策
Control and Decision

基于聚类簇结构特性的自适应综合采样法在入侵检测中的应用

刘金平， 周嘉铭， 刘先锋， 唐朝晖， 马天雨

引用本文:
刘金平, 周嘉铭, 刘先锋, 等. 基于聚类簇结构特性的自适应综合采样法在入侵检测中的应用[J]. 控制与决策, 2021, 36(8): 1920−1928.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2019.1672

基于聚类簇结构特性的自适应综合采样法在入侵检测中的应用

刘金平[1], 周嘉铭[1], 刘先锋[1†], 唐朝晖[2], 马天雨[1]

(1. 湖南师范大学 智能计算与语言信息处理湖南省重点实验室，长沙 410081；2. 中南大学 自动化学院，长沙 410083)

摘 要: 基于机器学习的网络入侵检测方法将恶意网络行为(入侵)检测转化为模式识别(分类)问题,因其适应性强、灵敏度高等优点,受到国内外广泛关注. 然而,现有的模式分类器往往假设数据集的分布是均衡的,而真实的网络环境中,入侵行为要远少于正常访问,这给网络入侵行为检测带来巨大挑战. 因此,提出一种基于聚类簇结构特性的综合采样法(CSbADASYN),通过挖掘少数类样本的内部结构对其进行自适应过采样,以获得样本分布结构特性保持的均衡数据样本,解决因数据不均衡带来的分类偏向. CSbADASYN先采用谱聚类方法对数据集中的少数类样本进行聚类分析,再根据所获得的聚类簇结构自适应插值,将获得样本分布结构保持的均衡样本用于分类器模型学习. 在经典的NSL-KDD和KDD99数据集上进行大量的验证性和对比性实验,结果表明,CSbADASYN能使传统分类器模型在不均衡数据集上的分类性能得到明显提升. 与传统的未经样本均衡处理和其他的带均衡处理的入侵检测方法相比,该方法能获得更低的误报率和漏报率.

关键词: 网络入侵检测；不均衡数据处理；分布结构保持；谱聚类；自适应综合采样法；过采样

中图分类号: TP273 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2019.1672 **开放科学(资源服务)标识码(OSID):**

引用格式: 刘金平,周嘉铭,刘先锋,等. 基于聚类簇结构特性的自适应综合采样法在入侵检测中的应用[J]. 控制与决策, 2021, 36(8): 1920-1928.

Toward intrusion detection via cluster structure-based adaptive synthetic sampling approach

LIU Jin-ping[1], *ZHOU Jia-ming*[1], *LIU Xian-feng*[1†], *TANG Zhao-hui*[2], *MA Tian-yu*[1]

(1. Hunan Provincial Key Laboratory of Intelligent Computing and Language Information Processing, Hunan Normal University，Changsha 410081，China；2. School of Automation，Central South University，Changsha 410083，China)

Abstract: Machine learning-based network intrusion detection identifies malicious network behaviors (intrusions) via pattern recognition (classification) technologies, which has attracted extensive attention due to its strong adaptability and high sensitivity. Existing pattern classifiers generally assume that the distributions of data sets are roughly balanced. However, in a real network environment, the number of intrusions is much less than the number of normal accessing. In this paper, a cluster structure-based adaptive synthetic sampling approach (CSbADASYN) is proposed, where minority classes are adaptively interpolated by mining the internal structure of minority-class samples to obtain the distributed characteristics-preserved balance samples for the detector training. The CSbADASYN adopts the spectral clustering method to cluster the minority-class samples in advance. Then, it makes an adaptive interpolation operation based on the achieved clusters to obtain balanced samples with distribution preserving characteristics for the classifier model learning. Extensive verification and comparative experiments are carried out on classic NSL-KDD and KDD99 datasets. Experimental results show that the CSbADASYN can significantly improve the classification performance of traditional classifier models on unbalanced datasets. Compared with other intrusion detection methods with equalization processing, the CSbADASYN can achieve lower false positive rate and false negative rate.

Keywords: intrusion detection；imbalanced data processing；distribution-structure persevering；spectral clustering；adaptive synthetic sampling approach；oversampling

收稿日期: 2019-11-28；**修回日期:** 2020-01-16.

基金项目: 国家自然科学基金项目(61971188)；湖南省自然科学基金项目(2018JJ3349)；湖南省教育厅优秀青年项目(19B364)；湖南省知识产权战略推进专项项目(2019F012K)；湖南省研究生科研创新项目(CX20190415).

责任编委: 阳春华.

†通讯作者. E-mail: xianfengliu_hunnu@163.com.

0 引　言

入侵检测作为一种主动的安全防护技术,能够在网络系统受到危害之前拦截和响应恶意的网络行为,近年来一直是信息安全领域的研究热点[1]. 区分网络事件(行为)是正常还是恶意是一种典型的模式分类问题,因而,研究者提出了多种基于机器学习的网络入侵检测方法,如决策树模型[2]、神经网络模型[3-4]、支持向量机模型[5]等. 还包括对这些模型的改进方法,例如:基于随机森林和支持向量机的网络入侵检测方法[6]、基于改进KNN的入侵检测方法[7]、蚁群算法和支持向量机结合的入侵检测方法[8].

在真实的网络环境中,入侵行为的数量往往远少于正常访问的数量,即正常网络访问数据和入侵访问数据具有严重的分布不均衡性. 例如在美国空军局域网上采集的网络连接数据中,普通用户对本地超级用户特权的非法访问远远少于正常访问[9];同时,不同的网络攻击行为之间也存在较大的数量差异,一些入侵行为(如拒绝服务攻击)的连接记录远多于其他的网络攻击(如提权攻击)的连接记录.

传统的机器学习模型在处理不均衡数据集时往往倾向于多数类而忽略少数类以获得更高的分类准确率,这会导致少数类难以有效区分. 因而,不平衡数据集的有效分类问题已经成为入侵检测领域的热门方向之一,其根本目标在于有效地提高少数类的分类准确率,从而提高入侵检测系统的性能[10]. 比如,Thomas[11]针对入侵数据的不均衡性,提出了一种基于数据决策融合的入侵检测技术,有效地提高了少数类检测的精确度.

目前,针对不均衡数据集进行处理的研究主要集中在算法层和数据层. 算法层面分为两大类:集成学习[12]和代价敏感学习[13]. 算法层处理方法的本质是对分类方法、准则进行改进,具有处理速度快的优点,但在精度提升上往往不如数据层处理方法. 数据层处理方法的核心是对数据进行过(欠)采样,从而改变多数类和少数类样本间的不均衡比,提高分类识别率. 目前,在不均衡数据的处理方法中,将数据重构与经典分类算法相结合已成为主流[14].

自适应综合过采样算法(adaptive synthetic sampling, ADASYN)[14]是一种面向数据层的不均衡数据处理方法,它根据少数类样本的概率分布r_i(i为少数样本合成数目的判定准则)自适应合成少数类样本,在难以分类的类别中合成更多的样本,使样本比例达到相对均衡的效果.

然而,原始的ADASYN算法在进行样本合成时只考虑了少数类样本周围的多数类样本的分布情况,没有考虑少数类样本特征之间的关联,未能充分利用其中的特征信息. 因而,如果能在充分考虑少数类样本的内部结构的基础上进行结构保持过采样,获得与原始样本空间分布结构一致的新样本进行分类器学习,从理论上将使分类器获得更好泛化性能.

本文针对网络入侵检测中因入侵行为作为少数类样本极易造成漏报或误报,且现有的面向不均衡数据集的ADASYN方法存在无法充分利用少数类样本结构信息的问题,提出一种基于聚类簇结构特性的综合采样法(CSbADASYN),实现结构保持的少数类样本过采样. CSbADASYN利用谱聚类方法将数据集分成若干个紧凑簇,有效获取稀疏类的空间结构;再以稀疏类的聚类簇中每个样本点与簇心的几何中心为单位进行样本插值;新插入的样本基于稀疏类的空间结构进行自适应插值,得到一个相对均衡的数据集,用于后续的入侵检测分类模型学习.

1 相关工作

本节简单回顾了ADASYN方法和谱聚类法的基本原理及其算法流程.

1.1 ADASYN

采样技术是对原始训练数据进行预处理,形成新的训练数据集,使新的数据集的多数类和少数类不再具有数量上的巨大差异. 从数据的增加或删减的角度,采样又分为过采样[15]和欠采样[16].

通过减少多类样本数量使数据集达到均衡为欠采样方法. 过采样从策略上分为随机式和启发式,随机采样只是简单随机删除和增加样本,而启发式采样则利用数据信息设计采样规则. SMOTE(synthetic minority oversampling technique)算法[17]是一种经典的过采样算法,该算法通过线性插值进行过采样. 然而,SMOTE算法在生成人工少数类样本过程中只是简单地在同类近邻样本间进行插值,对每个少数类样本合成数量相同的样本,并没有考虑到少数类样本周围多数类样本的分布情况,这使得该采样法存在一定的盲目性.

为弥补SMOTE算法进行过采样的盲目性,He等[14]提出了ADASYN,通过统计少数类样本周围多数类样本的情况,对少数类样本进行自适应的过采样,有效提高了少数类数据集的分类准确率. 研究表明,该方法能有效解决过采样中的盲目插值的问

题. ADASYN算法流程在文献[14]中有详细介绍.

1.2 谱聚类

谱聚类是一种经典的聚类算法,相较于其他聚类算法,谱聚类有两大优点: 1) 谱聚类只需要数据间的相似度矩阵,因此对数据分布的适应性较强,在数据较为稀疏时聚类效果也很好; 2) 谱聚类中蕴含了降维的思想,因此在处理高维数据聚类时的复杂度比传统聚类算法好. 本文选用谱聚类算法对少数类进行聚类处理,以获得其内部结构特性.

谱聚类的思想源于谱图划分问题,其本质是将聚类问题转化为无向图的多路划分问题. 将数据点视为无向图中的顶点,以无向图为基础,通过最优化某种划分准则,使得同类的点之间的点相似性较高,而不同类之间的点相似性较低. 不同的划分准则衍生出了不同的谱聚类算法[18].

谱聚类一般分为二路谱聚类算法和多路谱聚类算法. NJW算法[18]是一种经典的多路谱聚类算法,NJW谱聚类算法的核心思想是将数据点映射到特征空间后再进行聚类,从而得到原始数据的聚类结果. NJW算法是K-means算法的推广,在任意形状的数据上都具有较好的聚类效果,有着广泛的应用,因此,本文采用NJW谱聚类算法对数据进行处理.

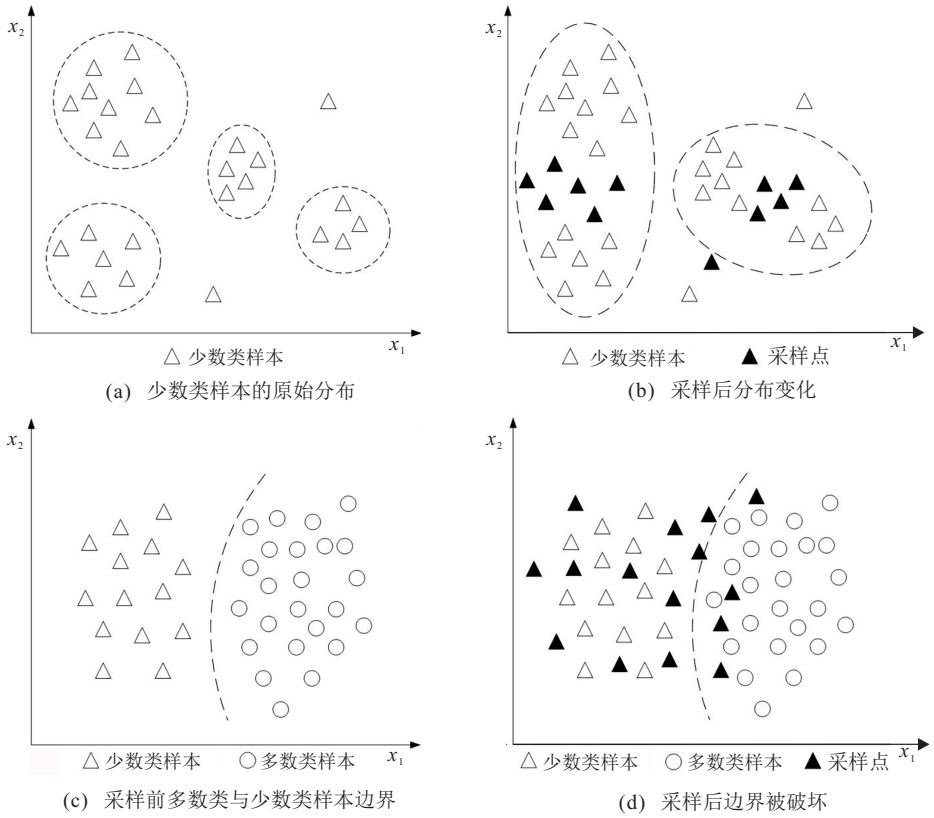
2 CSbADASYN

2.1 ADASYN过采样法局限性分析

ADASYN算法[14]的优点在于可以根据少数类样本的分布情况自适应合成样本,并且会在较难分类的地方合成更多样本,在较易分类的地方合成较少的样本. 然而,原始的ADASYN方法仍存在两点有待改进的地方.

1) 可能破坏少数类样本的原始分布.

ADASYN在计算插值数目时会考虑少数类样本周围的多数类分布情况,然而在少数类样本之间也存在特征信息的关联. 如果能充分利用少数类样本间的特征信息,再以此决定插值的数目和范围,将会进一步提高不均衡数据集的分类精度. 少数类样本间原有的分布情况可认为是一种特征信息,如图1(a)所示. 在进行插值前,少数类样本点群之间的分布信息本身就是一种有效样本结构特征信息.



(a) 少数类样本的原始分布　(b) 采样后分布变化

(c) 采样前多数类与少数类样本边界　(d) 采样后边界被破坏

图1 ADASYN采样前后样本间分布的变化

在采样时不考虑少数类样本之间的特征信息,则有可能出现如图1(b)所示的情况,即采样点出现在分布边缘之间,致使少数类样本之间的分布情况出现偏差甚至完全改变. 在数据量大、分类类别较多的情况中(如将过采样应用到入侵检测),将会破坏原始数据样本的空间分布特性,影响分类结果.

2) 可能影响多数类和少数类的边界.

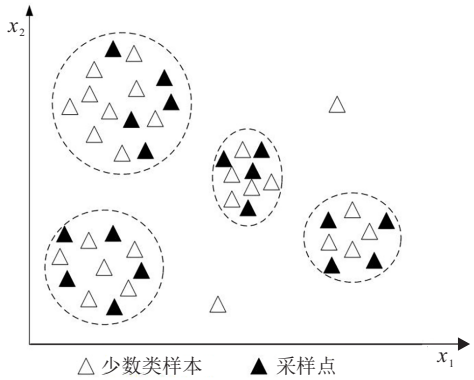
少数类样本与多数类样本的边界(如图1(c)所

示)作为一种重要的特征信息,与分类的精度息息相关. 如果对未经处理的样本点群直接采样,采样点可能会偏离少数类点群,如图1(d)所示. 通过ADASYN进行过采样后,新出现的采样点,极有可能破坏类别之间的边界,导致分类精度下降. 因此,在深入分析ADASYN的局限性后,本文提出一种聚类簇结构保持的自适应插值方法.该方法以少数类的空间聚类簇结构特性为基础,进行自适应过采样,以获得样本分布结构保持的均衡样本进行分类器学习.

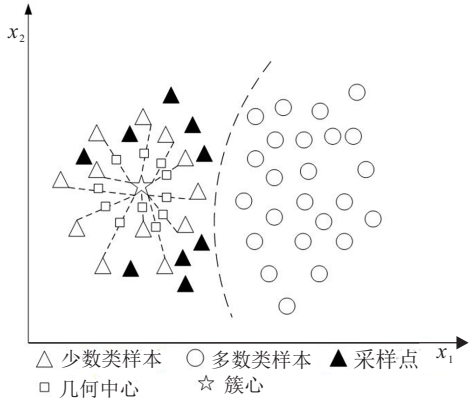
2.2　ADASYN的改进思路

针对ADASYN方法在不均衡数据处理中存在的问题,本文对ADASYN方法改进的思路立足于减少过采样对样本分布结构破坏,以获得分布结构保持的均衡样本,从而提高分类器的泛化能力.

1) 针对ADASYN破坏少数类样本原始分布的问题,本文采用一种基于图结构的聚类方法,即谱聚类算法,先对少数类样本进行聚类分析,获得其内部结构分布特性,再以各个聚类簇为基础进行过采样. 如图2(a)所示,使用对稀疏类聚类具有较高精度的谱聚类方法对少数类样本进行聚类处理,理论上可以有效地提取少数类样本间的特征信息. 同时,以稀疏类的紧聚类簇为基础对样本进行有约束地过采样,能够充分利用样本间的空间分布特性,维持少数类样本的原始分布特性.



(a)　采样后少数类样本维持原始分布



(b)　采样后维持少数类、多数类样本

图 2　以少数类簇的几何中心为单位采样

2) 对于ADASYN方法可能影响多数类和少数类样本边界的问题,本文以少数类样本点和聚类簇心的几何中心为单位,对过采样点进行约束,采用一种改进的过采样方法进行自适应插值.

图2(b)表示对谱聚类后某一个簇采用改进的方法进行过采样的示意图. 对于簇内的每一个样本点,找出其与簇心的几何中心,围绕几何中心进行插值,可有效限制插值的范围,理论上既可以利用样本点周围多数类样本的情况,也能够缓解采样后带来的类边界模糊,从而提高后续模式分类的准确性.

2.3　CSbADASYN算法流程

CSbADASYN首先计算出样本的不均衡度和总插值数,再利用谱聚类算法对少数类数据进行聚类,得到若干个聚类簇,根据每个聚类簇的少数类样本数目分配其插值数,最后以每个簇的簇心与样本点的几何中心为单位进行样本插值,得到一个相对均衡的数据集.

CSbADASYN具体的算法流程如下:

1) 计算不均衡度$d = m_s/m_l$,其中$d \in (0,1]$.

2) 计算合成少数样本数$G = (m_l - m_s)\beta$.其中: $\beta \in [0,1]$表示加入合成样本后的不均衡度, $\beta = 1$表示加入合成样本后多数类和少数类完全均衡,G等于少数类与多数类的差值.

3) 对少数类样本进行聚类,得到若干个簇C;计算簇之间的样本数比值,由此比值计算出每个簇的插值数G_i.

4) 对每个簇,找出每一个少数类样本x_i与簇心的几何中心o_i.

5) 对每个簇,以几何中心为单位进行插值. 对簇的每个几何中心o_i,找出它们在n维空间的K近邻,计算其比率$r_i = \Delta_i/K, i = 1,2,\ldots,m, r_i \in (0,1]$,其中$\Delta_i$是$o_i$的$K$近邻中多数类的数目.

6) 根据$\hat{r}_i = r_i \Big/ \sum_{i=1}^{m_s} r_i$正则化$r$. r_i为概率分布,$\sum \hat{r}_i = 1$,计算每个少数类样本周围多数类情况,并计算每个少数类样本周围多数类情况.

7) 根据每个几何中心o_i计算合成的样本数目g_i : $g_i = \hat{r}_i \times G_i$,其中$G$是合成样本的总数.

8) 在每个待合成少数类样本的几何中心周围K个邻居中选择1个少数类样本的几何中心,根据如下等式进行合成:

$$s_j = o_i + (o_{zi} - o_i)\lambda. \tag{1}$$

CSbADASYN在样本合成前加入了谱聚类. 谱

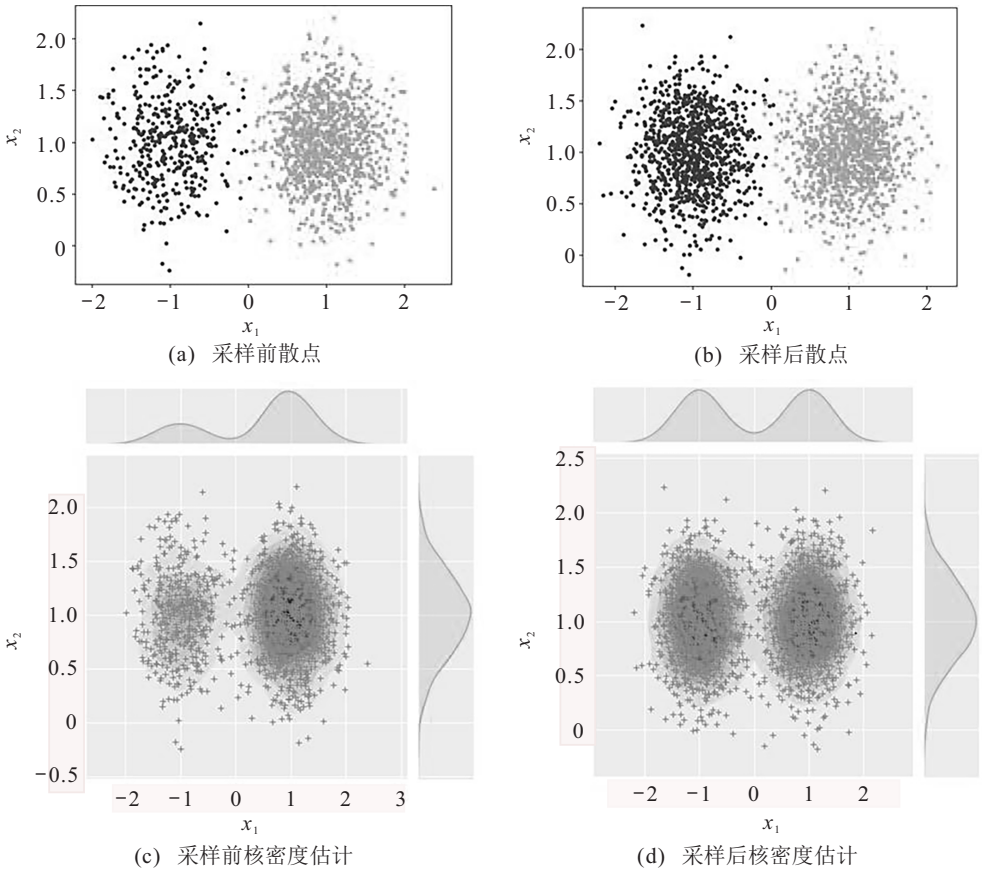
聚类根据数据的相似性矩阵的谱(特征值),在维度较少的数据聚类之前执行维数降低,可在小数据集上产生高质量的聚类. CSbADASYN在聚类阶段将少数类样本分为簇,其时间复杂度为$O(n^3)$;在插值阶段,时间复杂度为$O(kn)$,其中k是聚类的簇数,整个CSbADASYN的时间复杂度为$O(n^3)$.

2.4 数值仿真

首先采用两个简单的数值仿真实例来验证所提出的CSbADASYN方法的有效性,并通过验证过采样前后的概率密度来验证所提出方法对少数类过采样时的结构保持性.

实验1 两类别不均衡分布过采样实验.

生成2个相互独立且服从高斯分布的样本点集,共有1 400个样本点,每个样本点共有2维特征,中心分别为$(-1,1)$、$(1,1)$,点的数量分别为400和1 000,代表不均衡数据中的多数类与少数类. 将两类的初始方差均设置为0.38,其散点如图3(a)所示,图3(b)为其核密度估计.



(a) 采样前散点 (b) 采样后散点

(c) 采样前核密度估计 (d) 采样后核密度估计

图3 两类别样本散点图及核密度估计

采用本文提出的CSbADASYN对样本点集进行过采样,处理后的样本点集散点如图3(c)所示. 经CSbADASYN处理后,以$(-1,1)$为簇心的样本点群数目变为990. 由图3(c)可知,CSbADASYN对于样本点的采样有效地改善了类别之间的不均衡度.

为进一步验证CSbADASYN对原始样本空间分布结构的保持性,本文采用核密度估计(kernel density estimation)[19]的方法对图3(c)所示的样本点集进行非参数估计. 其概率密度分布如图3(d)所示,将图3(d)与图3(b)进行比较,可知经CSbADASYN处理后,样本点集的空间分布特性并未发生改变.

对样本方差的变化进行分析,分别使用原始的ADASYN以及本文提出的CSbADASYN进行处理. 实验10次取平均值,求出过采样后方差以及方差的变化率. 使用ADASYN处理后,方差为0.392,增大3.16%; 使用CSbADASYN处理后,方差为0.385,方差增大1.3%. 说明CSbADASYN在一定程度上降低了由过采样带来的样本方差变化,表明该方法在维持数据原有的空间分布结构上具有优越性.

实验2 多类别分布不均衡过采样实验.

生成5个服从高斯分布的点集,共1 600个3维样本点,其中以$(3,5)$为中心的样本点集有800个样本点,其余点集分别有200个样本点,通过选择不同的类别作为少数类进行数值仿真.

样本点集初始形态如图4(a)所示,中心为$(3,5)$的样本点群为多数类,其核密度估计如图4(d)所示.

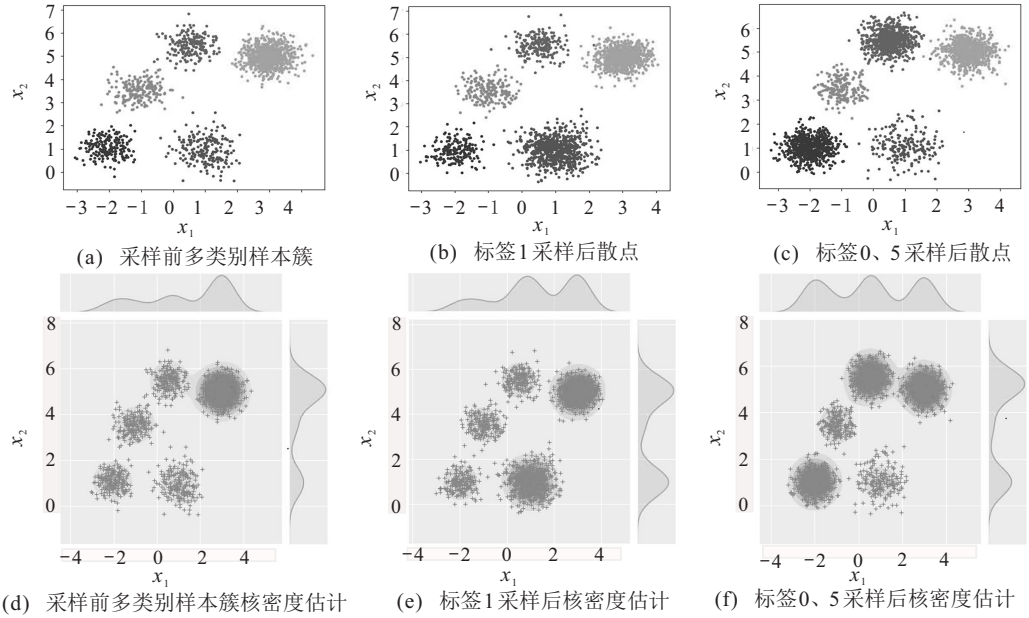


图4　多类别样本散点图及核密度估计

选择中心为(1, 1)、标签为1的样本点群作为少数类, 使用CSbADASYN对数据进行过采样, 图4(b)和图4(e)分别为采样后的散点和核密度估计. 由采样结果可知, CSbADASYN在处理多类别不均衡数据时性能稳定, 且采样点紧紧围绕在样本簇周围, 有效保持了数据的原始分布特性.

选择中心为(−2, 1)、(3, 5), 标签为0、5的两个样本点群作为少数类, 使用CSbADASYN对样本点群做过采样处理. 采样后的散点图以及核密度估计图分别如图4(c)和图4(f)所示. 在合并标签进行过采样时, CSbADASYN性能依然稳定, 且多数类与少数类边界未被模糊、 破坏, 表明本文所提出的CSbADASYN方法能有效完成对不均衡数据的过采样, 并且能够保证数据的结构稳定.

对于图4中的过采样的3类数据进行方差分析, 3类数据的初始方差均设置为0.4, 同时分别采用ADASYN、CSbADASYN进行过采样处理, 取10次独立实验的平均值.

ADASYN三组实验的方差值分别为0.412、0.423、0.420, 变化率分别为3.0%、 5.75%、 5.0%; CSbADASYN三组实验的方差值分别为0.412、0.423、0.420, 变化率分别为3.0%、5.75%、5.0%. 结果表明CSbADASYN对于样本结构维持的优越性在多类别不均衡数据的处理中依然保持稳定.

3　入侵检测实验

入侵检测实验主要包含两个部分:

1) 验证性实验: 在不均衡数据集上, 对本文提出的CSbADASYN进行有效性验证.

2) 对比性实验: 在KDD 99以及NSL-KDD数据集上, 与SMOTE、ADASYN算法进行对比, 分析本文算法在入侵检测中的实用性与优越性.

3.1　数据集介绍

1) KDD 99数据集.

KDD99是由美国国防部高级规划署在MIT林肯实验室模拟采集的网络连接数据集, 大约包含500万条网络连接数据, 分为训练集与测试集, 一共包含了4大类和39小类异常入侵类型以及正常连接. 其标识类型如下[20]: Normal(正常记录)、Dos(拒绝服务攻击)、Probe(监视和其他探测活动)、R2L(来自远程机器的非法访问)、U2R(普通用户对本地超级用户特权的非法访问).

2) NSL-KDD数据集.

NSL-KDD是KDD99数据集的改进. 与KDD99数据集相比, NSL-KDD数据集[21]的训练集中不包含冗余记录, 所以分类器不会偏向更频繁的记录; NSL-KDD数据集的测试集中删除了大量重复记录; NSL-KDD中来自每个难度级别组的所选记录的数量与原始KDD数据集中的记录的百分比成反比; NSL-KDD训练和测试中的记录数量设置是合理的, 这使得在整套实验上运行实验成本低廉.

3.2　评价标准

分类后的数据分为4种: 正确分类的多数类、正确分类的少数类以及错误分类的多数类、错误分类的少数类, 分别表示为TN、TP、FP、FN. 在不均衡数据集的分类评价中, 少数类数据的性能检测评价的意义比总体性能评价的意义更大. 因此, 本文采用*G*-

mean、F-measure[22]、AUC作为评价标准.

G-mean的取值取决于多数类精确度和少数类精确度乘积的平方根. 只有当两者的值都较大时,即多数类与少数类的分类精确度都较为高时,G-mean的取值才会大. G-mean的计算公式如下:

$$G\text{-mean}=\sqrt{\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}\times\frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}}. \tag{2}$$

F-measure计算公式如下:

$$F\text{-measure}=\frac{(1+\beta^2)\cdot\mathrm{recall}\cdot\mathrm{precision}}{\beta^2\cdot\mathrm{recall}+\mathrm{precision}}. \tag{3}$$

F-measure是精确率和召回率的组合. 当两者同时趋于最大值时,F-measure的值趋于最大值. F-measure的值越大,说明不均衡数据的分类效果越好.

ROC曲线图的纵轴是TPrate,横轴是FPrate. 曲线越靠近左上角表示分类器性能越好. 由于曲线不能定量地对分类器的性能进行评价,常用曲线下面积AUC作为评价方法,值越大,分类器的性能越好.

在对比性实验中,本文还将采用准确率(AR)和漏报率(MR)两个指标作为评价标准,有

$$\mathrm{AR}=\frac{\mathrm{TP}+\mathrm{TN}}{P+N}, \tag{4}$$

$$\mathrm{MR}=1-\frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}. \tag{5}$$

其中:P为正例个数,N为负例个数. 准确率能客观地评估一个算法的综合性能,而漏报率能有效地评估算法的泛化性[23].

3.3 实验结果及分析

3.3.1 验证性实验

在验证性实验中,本文先选用了UCI数据集中的pima数据集进行少数类过采样和分类实验验证. 该数据集有768个样本,其中多数类样本数目为500,少数类样本数目为268. 比较未处理、ADASYN算法处理以及CSbADASYN算法处理后的pima数据集,在多个分类器下的分类结果如表1所示.

表1 验证性实验结果

classifier	algorithm	F-measure	AUC	G-mean
SVM	未处理	0.668 6	0.705 2	0.656 5
	ADASYN	0.786	0.779 3	0.778 8
	CSbADASYN	0.818 1	0.928 1	0.815 2
RF	未处理	0.725 4	0.706	0.726 5
	ADASYN	0.761 3	0.781 7	0.780 8
	CSbADASYN	0.782 1	0.800 5	0.800 4
MLP	未处理	0.695 1	0.705 4	0.715 5
	ADASYN	0.760 9	0.748 1	0.747
	CSbADASYN	0.776 2	0.744 6	0.739 3

由表1中的实验结果可知,经CSbADASYN过采样方法处理后,使用随机森林(RF)、支持向量机(SVM)、以及多层感知机(MLP)对数据集分类,相较于ADASYN算法,F-measure、G-mean以及AUC值分别提高了4.08%、4.67%、7.19%,表明采用CSbADASYN对不平衡数据集处理后,分类器的分类效果、对于少数类样本的分类精度均有提高.

3.3.2 对比性实验

1) NSL-KDD: 5分类.

在NSL-KDD中,训练和测试中的记录数量设置较为符合传统的分类模型训练和性能验证,使得在整套实验上运行成本低廉而无需随机选择一小部分. 基于此特性,本文在NSL-KDD的4种入侵模式中进行对比性入侵检测实验,比较经过CSbADASYN处理后,对RF、SVM以及MLP分类器的准确率和漏报率的影响. 实验结果如表2所示,其中CSbADASYN+RF,CSbADASYN+SVM和CSbADASYN+MLP分别代表所提出的CSbADASYN过采样法与相应的分类器模型相结合进行入侵检测的实验结果.

表2 对比性实验结果(NSL-KDD)

算法	评价指标	入侵类别			
		DoS	Probe	R2L	U2R
RF	准确率	0.912 3	0.927 4	0.935 3	0.923 4
	漏报率	0.626 3	0.602 3	0.604 1	0.610 2
CSbADASYN+RF	准确率	0.924 1	0.915 4	0.944 3	0.944 5
	漏报率	0.432 6	0.394 5	0.521 2	0.476 2
SVM	准确率	0.914 6	0.914 5	0.902 3	0.904 1
	漏报率	0.261 6	0.221 3	0.216 4	0.194 3
CSbADASYN+SVM	准确率	0.932 4	0.934 1	0.925 4	0.926 2
	漏报率	0.223 7	0.196 4	0.201 3	0.181 2
MLP	准确率	0.915 2	0.909 3	0.901 2	0.910 3
	漏报率	0.301 2	0.361 2	0.354 1	0.367 2
CSbADASYN+MLP	准确率	0.932 1	0.921 2	0.931 4	0.947 4
	漏报率	0.296 7	0.356 1	0.305 2	0.336 3

融合CSbADASYN的入侵检测模型在准确率和漏报率的表现都有提升. 其中,在RF对照实验中,平均准确率提高了1.3%,平均漏报率降低了8.83%;在SVM对照实验中,平均准确率提高了2%,平均漏报率降低了9.8%;在MLP对照实验中,平均准确率提高了2.33%,平均漏报率降低了5.77%. 说明经CSbADASYN处理后,数据的不均衡得到有效处理,使分类器的精度提升,从而有效提升了入侵检测模型的准确性,降低了模型的漏报率.

在入侵类别中,R2L以及U2R属于少数类. 本文方法的加入有效提升了它们的准确率,验证了本文所提出方法对于检测少数入侵类样本的优越性.

2) KDD 99(10 %): 10分类.

NSL-KDD数据集是经过人工筛选的数据集,数据集中各种类别的数量差异性与真实的网络环境差异较大. 因此,本文选用数据更贴近真实网络KDD 99再进行小类别的入侵检测(分类)测试,以此检验经CSbADASYN方法处理后的模型对于少数类入侵模式的分类能力.

实验选用back、pod、teardrop、ipsweep、nmap,portsweep、satan、guess_password、warezclient、butter_overflow等10个攻击模式作为标签,探究是否加入本文方法,对少数类入侵模式检测的影响. 实验结果如表3所示.

表 3　对比性实验结果(KDD 99)

算法	评价指标	入侵类别									
		back	pod	teardrop	ipsweep	nmap	portsweep	satan	guess_p	warezclient	butter_o
RF	*F*-measure	0.632 1	0.724 1	0.901 5	0.913 2	0.931 4	0.970 5	0.931 4	0.846 2	0.923 1	0.903 8
	G-mean	0.725 2	0.731 3	0.915 4	0.906 4	0.921	0.920 3	0.920 3	0.831 4	0.946 2	0.932 4
	AUC	0.722 6	0.755 6	0.902 3	0.905 5	0.922 2	0.921 3	0.932 1	0.842 2	0.912 3	0.925 4
CSbADASYN +RF	*F*-measure	0.691 5	0.853 1	0.931 4	0.921 4	0.946 4	0.971 4	0.936 3	0.886 4	0.933 3	0.913
	G-mean	0.731 3	0.801 4	0.935 2	0.951 8	0.937 4	0.946 3	0.921 3	0.872 5	0.942 1	0.943 2
	AUC	0.702 1	0.752 3	0.915 6	0.903 3	0.924 1	0.923 2	0.914 8	0.882 2	0.931 2	0.913 2
SVM	*F*-measure	0.572 7	0.792 3	0.942 3	0.961 4	0.926 4	0.916 4	0.932 4	0.903 6	0.932 1	0.901 2
	G-mean	0.629	0.781 5	0.921 5	0.942 3	0.913 2	0.923 6	0.946	0.916 2	0.942 5	0.932 3
	AUC	0.600 2	0.732 3	0.901 3	0.923 2	0.902 2	0.921 4	0.932 5	0.912 1	0.933 3	0.932 1
CSbADASYN +SVM	*F*-measure	0.591 5	0.823 2	0.925 7	0.914 2	0.935 4	0.932 1	0.933 3	0.932 6	0.942 3	0.912 8
	G-mean	0.673 2	0.842 5	0.913 2	0.915 6	0.946 3	0.951	0.932	0.943	0.932	0.936 2
	AUC	0.653 2	0.832 1	0.903 6	0.903 1	0.913 1	0.94				
MLP	*F*-measure	0.592 3	0.7451	0.923 6	0.925 2	0.913 5	0.914 6	0.934 9	0.893 2	0.915 4	0.920 3
	G-mean	0.621 4	0.751 3	0.955 1	0.934 5	0.912 3	0.910 3	0.914 6	0.876 2	0.936 4	0.943
	AUC	0.623 2	0.922 3	0.922 3	0.942 1	0.903 2	0.903 3	0.900 1	0.885 6	0.921 4	0.931 5
CSbADASYN +MLP	*F*-measure	0.617 3	0.826 5	0.956 5	0.954 2	0.924 7	0.931	0.934 2	0.889 3	0.953 6	0.931 2
	G-mean	0.651 3	0.792 4	0.961 4	0.943 1	0.931 5	0.926 1	0.962	0.903 4	0.946 3	0.906 2
	AUC	0.666 2	0.801 4	0.953 1	0.923 2	0.923 2	0.901 2	0.942 3	0.913 2	0.921 3	0.911 2

由实验结果可知,加入本文提出的CSbADASYN过采样方法后,RF、SVM和MLP的*G*-mean分别提高2.3 %、3.0 %、1.7 %. *G*-mean是保持多数类、少数类分类精度均衡的情况下最大化两类的精度,即只有在多数类和少数类的分类精度同时都高的情况下,*G*-mean才会大,说明本文方法使得不均衡数据的整体分类性能提高. RF、SVM和MLP的*F*-measure分别提高3.8 %、1.8 %、2.7 %. *F*-measure提升说明数据集在分类过程中,在整体数值的分类精度得到提升的情况下,少数类样本的分类精度获得提升的程度更高. RF、SVM和MLP的AUC分别提高了2.4 %、1.9 %、2.1 %,说明经本文方法处理后,分类器的整体性能得到提升.

实验结果还表明,在使用更大的数据集、更小的分类标签时,本文提出的方法的性能稳定. 每个少数类攻击模式(在本例中为“back”“guess_password”“tear_drop”“warez_client”等数量较少的攻击模式)依然保持有较高的*F*-measure、*G*-mean以及AUC. 充分表明了本文提出的方法在入侵检测中对于少数类入侵模式具有较好的检测能力.

4　结　论

本文提出了一种基于少数类聚类簇空间结构分布特性的自适应综合采样法(CSbADASYN),对非均衡数据集中的少数类样本进行过采样,以获得相对均衡的数据样本用于分类模型学习. CSbADASYN采用谱聚类算法将少数类样本分成若干个簇,再以簇为单位对少数类样本进行以聚类簇几何中心位置为基础的自适应插值,以此改善数据的均衡度. 数值仿真实验以及分类实验结果表明,经CSbADASYN过采样法获得的新训练数据集能有效保持样本的空间分布特性,因而能使传统分类器模型在不均衡数据集上的分类性能得到明显提升,利于不均衡数据的处理. 将CSbADASYN与传统的分类器相结合构成新的入侵检测模型,在NSL-KDD、KDD 99数据集进行大量的验证性和对比性实验,结果表明:融合CSbADASYN的入侵检测模型能有效检测出各种类型的入侵类型,同时有效降低了入侵检测的误报率和漏报率.

参考文献(References)

[1]　刘金平, 张五霞, 唐朝晖, 等. 基于模糊粗糙集属性约简与GMM-LDA最优聚类簇特征学习的自适应网络入侵检测[J]. 控制与决策, 2019, 34(2): 243-251.

(Liu J P, Zhang W X, Tang Z H, et al. Adaptive network intrusion detection based on fuzzy rough set-based attribute reduction and GMM-LDA-based optimal cluster feature learning [J]. Control and Decision, 2019, 34(2): 243-251.)

[2] Sahu S, Mehtre B M, Sahu S, et al. Network intrusion detection system using J48 Decision tree[C]. International Conference on Advances in Computing, Communications and Informatics (ICACCI). Kochi: IEEE, 2015: 2023-2026.

[3] Hodo E, Bellekens X, Hamilton A, et al. Threat analysis of IoT networks using artificial neural network intrusion detection system[J]. Tetrahedron Letters, 2017, 42(39): 6865-6867.

[4] Amini M, Jalili R, Shahriari H R. RT-UNNID: A practical solution to real-time network-based intrusion detection using unsupervised neural networks[J]. Computers & Security, 2006, 25(6): 459-468.

[5] Duan X H, Liu Z G, Liu H. Network intrusion detection by rough set and least squares support vector machine[C]. International Conference on Signal Processing Systems. Dalian: IEEE, 2010: 564-566.

[6] Chang Y, Li W, Yang Z. Network intrusion detection based on random forest and support vector machine[C]. 2017 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC). Zhuhai: IEEE, 2017: 635-638.

[7] Jaiswal S, Saxena K, Mishra A, et al. A KNN-ACO approach for intrusion detection using KDDCUP'99 dataset[C]. The 3rd International Conference on Computing for Sustainable Global Development (INDIACom). Shanghai: IEEE, 2016: 628-633.

[8] Cai X, Yu F. A wavelet transform based support vector machine ensemble algorithm and its application in network intrusion detection[J]. International Journal of Security & Its Applications, 2015, 9(4): 307-316.

[9] Siddique K, Akhtar Z, Khan F A, et al. KDD cup 99 data sets: A perspective on the role of data sets in network intrusion detection research[J]. Computer, 2019, 52(2): 41-51.

[10] Zhou P, Hu X, Li P, et al. Online feature selection for high-dimensional class-imbalanced data[J]. Knowledge-Based Systems, 2017, 136(1): 187-199.

[11] Thomas C. Improving intrusion detection for imbalanced network traffic[J]. Security and Communication Networks, 2013, 6(3): 309-324.

[12] Qian Y, Liang Y, Li M, et al. A resampling ensemble algorithm for classification of imbalance problems[J]. Neurocomputing, 2014, 143: 57-67.

[13] Liu M, Xu C, Luo Y, et al. Cost-sensitive feature selection by optimizing F-measures[J]. IEEE Transactions on Image Processing, 2017, 27(3): 1323-1335.

[14] He H, Bai Y, Garcia E A, et al. ADASYN: Adaptive synthetic sampling approach for imbalanced learning[C]. Proceedings of International Joint Conference on Neural Networks (IJCNN). Hong Kong: IEEE, 2008: 1322-1328.

[15] Ahmad J, Javed F, Hayat M. Intelligent computational model for classification of sub-Golgi protein using oversampling and fisher feature selection methods[J]. Artificial Intelligence in Medicine, 2017, 78: 14-22.

[16] Lin W C, Tsai C F, Hu Y H, et al. Clustering-based undersampling in class-imbalanced data[J]. Information Sciences, 2017, 409/410: 17-26.

[17] Chawla N V, Bowyer K W, Hall L O, et al. SMOTE: Synthetic minority over-sampling technique[J]. Journal of Artificial Intelligence Research, 2002, 16(1): 321-357.

[18] Ng A Y, Jordan M I, Weiss Y. On spectral clustering: Analysis and an algorithm[C]. Advances in Neural Information Processing Systems. San Francisco: Researchgate, 2002: 849-856.

[19] Aitchison J, Lauder I J. Kernel density estimation for compositional data[J]. Applied Statistics, 2018, 34(2): 129-137.

[20] Tavallaee M, Bagheri E, Lu W, et al. A detailed analysis of the KDD CUP 99 data set [C]. 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications. Ottawa: IEEE, 2009: 1-6.

[21] Deshmukh D H, Ghorpade T. Improving classification using preprocessing and machine learning algorithms on NSL-KDD dataset[C]. 2015 International Conference on Communication. London: IEEE, 2015: 1-6.

[22] Dembcz K, Waegeman W, Cheng W, et al. An exact algorithm for F-measure maximization[C]. Advances in Neural Information Processing Systems. New York: Curran Associates Inc, 2011: 1404-1412.

[23] 刘金平, 何捷舟, 马天雨, 等. 基于KELM选择性集成的复杂网络环境入侵检测[J]. 电子学报, 2019, 47(5): 1070-1078.
(Liu J P, He J Z, Ma T Y, et al. Selective ensemble of KELM-based complex network intrsion detection[J]. Chinese Journal of Electronics, 2019, 47(5): 1070-1078.)

作者简介

刘金平 (1983−), 男, 副教授, 从事复杂工业过程智能监测、建模与优化控制等研究, E-mail: ljp202518@163.com;

周嘉铭 (1996−), 男, 硕士生, 从事模式识别、数据挖掘的研究, E-mail: zhoujiaming_hunnu@163.com;

刘先锋 (1964−), 男, 教授, 从事数据库理论与应用、数据挖掘等研究, E-mail: xianfengliu_hunnu@163.com;

唐朝晖 (1965−), 男, 教授, 从事复杂工业系统的建模与优化控制、信息处理等研究, E-mail：zhtang@csu.edu.cn;

马天雨 (1978−), 男, 讲师, 从事复杂工业过程、智能控制的研究, E-mail：mty@hunnu.edu.cn.

(责任编辑：孙艺红)