DOI:10.13718/j.cnki.xsxb.2018.12.016

基于密度和最优聚类数的入侵检测方法①

邹臣嵩[1], 杨 宇[2]

广东松山职业技术学院 1. 电气工程系; 2. 机械工程系, 广东 韶关 512126

摘要: 针对聚类算法在入侵检测应用中存在的参数预设、聚类有效性评价、未知攻击类型检测等问题, 提出了一种基于密度和最优聚类数的改进算法, 根据样本的分布情况启发式地确定初始聚类中心, 从样本的几何结构角度提出一种新的内部评价指标, 给出了最优聚类数确定方法, 在此基础上, 设计了一个增量式的入侵检测模型, 实现了聚类中心和聚类数目的动态调整. 实验结果表明, 与 K-means 及其他两种改进聚类算法相比, 新算法收敛速度更快、聚类准确率更高, 能够对未知网络行为进行有效聚类, 具有较好的入侵检测效果.

关 键 词: 聚类算法; 最优聚类数; 入侵检测; 有效性评价; 密度聚类

中图分类号: TP393.08 **文献标志码**: A **文章编号**: 1000-5471(2018)12-0091-09

基于聚类分析的入侵检测技术可以不依赖标记样本就能实现对入侵行为的有效检测[1]. 由于入侵检测系统获取的网络行为记录具有很强的随机性, 其聚类数目和聚类结果形状事先难以确定, 故无法直接使用传统的 K-means 算法[2], K-medoids 算法[3]和模糊 C 均值算法[4]. 对于这种无先验知识的样本, 通常的解决方法是: 借助专家经验预设一个聚类基数 k, 或给定一个阈值范围[5], 再通过对 k 值的增减调整完成对样本的训练, 经多次计算对比后, 得到相对理想的聚类结果. 鉴于入侵检测数据的维度高、数量大, 此类算法在具体应用时运算耗时长, 整体效率不高. 此外, 由于引入了专家经验和预设参数, 必然导致聚类结果中存在部分主观性因素, 算法的稳定性无法得到保障. 因此, 如何确定网络行为记录集的最优聚类数目或阈值范围是入侵检测技术亟需解决的重要问题.

1 相关算法研究现状

在聚类算法改进方面, Breunig 等[6]将基于密度的聚类算法 OPTICS 与异常检测结合在一起, 提出了基于局部异常的概念及相应的异常检测方法, 该方法通过使用局部异常因子 LOF(Local Outlier Factor)来度量对象的差异程度, 是一种基于相对密度的异常检测方法. 蒋盛益等[7]用幂平均取代 LOF 定义中的简单算术平均, 并用平均可达距离代替局部可达密度, 将局部异常因子的定义推广到了更一般的情况. 翟东海等[8]提出了最大距离法选取初始簇中心的聚类算法, 根据簇内距离和最小思想重新设计了迭代过程中的簇中心计算方法. 熊忠阳等[9]在相对密度概念的基础上对最大距离法进行了改进, 使得初始聚类中心具有良好分散性和代表性, 解决了最大距离法对非球状数据检测效果不佳的缺陷, 但是在具体使用中, 其密度参数需要人工预设, 自适应性较差. 相比其他聚类算法, AP 算法可以更快地处理大规模数据, 同时避免聚类中心的初始选择, 但对于非团状的数据集, AP 算法倾向于产生较多的局部聚类, 聚类数目与实际值有所偏差[10-11]. 此外, 由于 AP 算法的聚类数目受到参考度 p 的影响, 而 p 值的选取目前尚无成熟的理论依据,

① 收稿日期: 2018-05-15

基金项目: 广东高校省级重大科研项目(2017GkQNCX033); 韶关市科技计划项目(2017CX/K055); 广东松山职业技术学院重点科技项目(2018KJZD001); 广东大学生科技创新培养专项(pdjh2015a0715).

作者简介: 邹臣嵩(1980-), 男, 讲师, 硕士, 主要从事数据挖掘与网络安全的研究.

这样就导致了该算法大多以局部最优或近似全局最优作为最终结果，研究表明，选取相似度的中位数作为 p 值得到的聚类数往往大于或等于正确类数[12]，这虽然具备一定的参考价值，但精度依旧不足，因此需要对聚类数的合理性进一步分析与评价，以达到更理想的效果.

内部评价指标未使用样本的先验知识，通过计算簇内相似度、簇间相似度来评价聚类效果优劣，是发现数据集最佳聚类数的常用办法[13-14]，常用的基于样本几何结构的内部指标有 IGP 指标[15]、DBI 指标[16]、Sil 指标[17]等. 为了解决聚类结果的有效性评价，Xie 和 Beni 结合样本集的几何特征，将簇内紧致度和簇间分离度的比值作为聚类有效性指标[18]. Kwon 在 Xie-Beni 指标的基础上，引入了惩罚函数，有效地解决了当 $c \to n$ 时，Xie-Beni 指标递减趋于 0 的问题，提出了 S. H. Kwon 有效性指标[19]. 任敏[20]在 Xie-Beni 指标的基础上，提出了一个新的模糊聚类有效性指标，结合穷举搜索策略，以基于局部密度的初始聚类中心选择算法获取的最大聚类数目作为搜索上界，求解出数据集的最佳聚类数目.

上述文献虽然在一定程度上解决了聚类算法在入侵检测应用中存在的问题，但仍存在参数选取敏感、过度依赖先验知识，聚类有效性检验效果不理想等不足. 针对这些问题，本文从聚类算法改进和最优聚类数确定两个方面对入侵检测技术中的聚类问题进行了研究，提出了基于密度和最优聚类数的入侵检测方法(Density and Optimal Clustering Number，以下简称 DOCN).

2 基于密度和最优聚类数的入侵检测方法

DOCN 由两部分组成，结构如下：

1) 聚类：根据样本的分布情况自动匹配密度参数，启发式地构造高密度点集合，进而得到较优初始聚类中心，不断更新簇中心，直至收敛，最终完成聚类.

2) 确定最优聚类数与最优聚类结果：根据簇内与簇间的几何特征，提出了一种新的内部评价指标，通过对样本的聚类结果进行优劣评价，确定最优聚类数与最优聚类结果.

需要特别指出的是，为了缩小聚类数的搜索范围，本文使用 AP[21]算法对网络行为记录集进行初步聚类，由于 AP 算法中参考度 p 值大小与最终的聚类中心数目呈现正相关的关系，故 p 的取值范围决定了聚类数的空间大小. 因此，在具体应用中，本文通过调节参数 p，获取一系列具有差异性的聚类数作为参考值，随后结合改进的密度算法完成二次聚类，最后利用新评价指标从样本几何的角度对二次聚类的结果进行分析与评价，进而确定最佳聚类数.

2.1 基本概念与公式

设 $X=\{X_1, X_2, \cdots, X_i, \cdots, X_n\}$ 为含有 n 个数据对象的网络行为记录集，每个样本含有 q 个特征. 现将该集合划分为 k 个簇，即 $X=\{\mathrm{Cluster}_1, \mathrm{Cluster}_2, \cdots, \mathrm{Cluster}_k\}$，每簇含样本 m 个，簇中心所构成的集合 $C=\{C_1, C_2, \cdots, C_k\}(k<n)$. X_{ij} 是第 i 簇的第 j 个数据对象，C_i 是第 i 簇的中心.

2.1.1 基于密度的改进聚类算法

最优聚类数的获取除了和有效性指标本身有关外，还与所采用的聚类算法密不可分. 鉴于密度算法具有发现任意形状空间聚类的优点，本文对密度算法从初始聚类中心选择和簇中心更新两个方面进行了改进，基本思路如下：

在初始聚类中心阶段，首先计算样本集中各数据对象的密度和样本集的平均密度，将密度大于样本平均密度的数据对象存储至高密度点集合，再将距离高密度中心最远的高密度点 C_1 作为首个初始聚类中心存储至聚类中心集合 C 中，然后将距离高密度中心和 C_1 乘积最大的高密度点 C_2 存储至 C 中，以此类推，得到全部初始聚类中心，即 $C=\{C_1, C_2, \cdots, C_k\}$.

在更新簇中心阶段，首先依据初始聚类中心完成首次聚类，接着选取与簇内其他样本距离之和最小的数据对象作为簇中心，进而构建新的簇中心集合 $C'=\{C'_1, C'_2, \cdots, C'_k\}$，再将样本集 X 中其他数据对象按最小距离划分到相应簇中，重复迭代过程，直至准则函数收敛，最终完成聚类. 算法的相关定义和公式如下：

定义 1 空间两点间的欧氏距离定义为

$$d(X_i, X_j)=\sqrt{\sum(X_i-X_j)^2} \tag{1}$$

其中：$i=1, 2, \cdots, n$；$j=1, 2, \cdots, n$

定义 2　集合 X 中任意两点间的平均距离定义为

$$d_{\text{avg}} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} d(X_i, X_j)}{A_n^2} \tag{2}$$

其中：A_n^2 表示从样本集中任意选取两个样本的所有排列.

定义 3　样本 X_i 的密度定义为以 X_i 为圆心，以 d_{avg} 为半径的圆内(含圆上)所包含数据对象的个数，即当 X_i 与 X_j 的距离小于或等于样本集的平均距离时，计数器加 1，即：

$$(X_i)_{\text{dens}} = \sum_{j=1}^{n} \text{count}(d(X_i, X_j) \leqslant d_{\text{avg}}) \tag{3}$$

其中：$i=1, 2, \cdots, n$；$j=1, 2, \cdots, n$

定义 4　样本集 X 的平均密度定义为各样本的密度之和除以样本总数，即：

$$D_{\text{avg}} = \frac{\sum_{i=1}^{n}(X_i)_{\text{dens}}}{n} \tag{4}$$

定义 5　高密度点集合 D 定义为密度高于样本集平均密度的数据对象组成的集合.

定义 6　高密度中心是高密度点集合 D 的均值，即：

$$D_{\text{center}} = \frac{D}{|D|} \tag{5}$$

定义 7　样本 X_i 与簇内其他数据对象的距离之和(S) 定义为

$$S\_X(i) = \sum_{j=1}^{m} d(X_i, X_j) \tag{6}$$

其中：$X_i \in \text{Cluster}_t$，$X_j \in \text{Cluster}_t$，$t=1, 2, \cdots k$

定义 8　簇内距离和矩阵定义为

$$S\_\text{Cluster}(t) = \begin{bmatrix} S\_X(1) \\ S\_X(2) \\ \cdots \\ S\_X(m) \end{bmatrix} \tag{7}$$

其中：$t=1, 2, \cdots, k$

定义 9　在簇中心更新过程中，将与簇内其他样本距离之和最小的数据对象 X_i 作为该簇的中心，X_i 满足以下条件

$$S\_X(i) = \min(S\_\text{Cluster}(t)) \tag{8}$$

其中：$t=1, 2, \cdots, k$

2.1.2　聚类有效性评价指标

聚类有效性评价指标是对聚类结果进行优劣判断的依据，通过比较指标值可以确定最佳聚类划分和最优聚类数，在对无先验知识样本的聚类结果进行评价时，通常将“簇内凝聚，簇间分离”作为内部评价的重要标准，从簇内凝聚的角度考虑，我们希望样本之间的距离越近越好，从簇间分离的角度考虑，希望各簇之间的距离越远越好. 在对数据集进行聚类划分过程中，如果数据集的簇间距离大且簇内距离小，说明此时的簇内凝聚度和簇间分离度较为理想，为平衡二者关系，本文用表达式：$\frac{\text{簇间距离}}{\text{簇内距离}} - \frac{\text{簇内距离}}{\text{簇间距离}}$ 作为内部评价指标来评价无先验知识样本的聚类有效性，由表达式可知，数据集的簇间距离与簇内距离的比值越大，该指标值也越大，当指标值达到最大时，意味着此时簇间与簇内的划分最为理想，说明了此时的聚类结果和聚类数目最优. 评价指标的具体定义和公式如下：

定义 10　样本 X_i 的簇内距离为该样本到簇内其他数据对象的平均距离，即：

$$w_i = \frac{1}{m-1}\sum_{j=1,\ j\neq i}^{m} d(X_i, X_j) \tag{9}$$

其中：$X_i \in \text{Cluster}_t$，$X_j \in \text{Cluster}_t$，$t=1,2,\cdots k$

定义11 样本X_i的簇间距离为该样本到其他各簇中心距离的最小值，即：

$$b_i = \min_{1<j\leqslant k,\ X_i \notin \text{Cluster}_j} (d(X_i, C_j)) \tag{10}$$

其中：$X_i \in \text{Cluster}_t$，$C_j \in \text{Cluster}_j$，C_j表示第j簇的簇中心，$t=1,2,\cdots k$，$j=1,2,\cdots k$，$t \neq j$

定义12 样本X_i的聚类评价指标(Cluster Evaluation Index，简称CEI)定义为

$$CEI_i = \frac{b_i}{w_i} - \frac{w_i}{b_i} \tag{11}$$

其中：$t=1,2,\cdots,k$

CEI_i指标反映了样本X_i的聚类有效性，CEI_i越大，意味着样本X_i的聚类质量越好. 本文使用全部样本的CEI平均值来评价其整体聚类效果，该平均值越大，说明样本集的整体聚类质量越好.

定义13 样本集X聚为k类时的CEI平均值为

$$CEI(k) = \frac{1}{n}\sum_{i=1}^{n} CEI_i \tag{12}$$

定义14 最优聚类数k定义为$CEI(k)$取最大值时的聚类数目，即：

$$k_{\text{opt}} = \arg\max_{k_{\min}\leqslant k\leqslant k_{\max}} \{CEI(k)\} \tag{13}$$

其中：$k_{\min}$和$k_{\max}$由改进的AP算法[21]得出.

2.2 算法描述

DOCN可细分为5个环节：

2.2.1 K值范围的获取

提取网络行为记录集中的字符型数据，对其进行独热编码后再与原样本中的连续型数据重新组合，构建新的样本集X，设置AP算法中的参数$p_{\max}$为相似度的中位数，$p_{\min} = p_{\max}/2$，执行算法后，得到X的聚类数范围$[k_{\min}, k_{\max}]$，令k初值为$k_{\min}$.

2.2.2 选择初始聚类中心

a) 根据式(1)—(3)计算样本集X中每条行为记录的密度，根据式(4),(5)得到高密度点集合D和高密度中心D_{center}；

b) 根据式(1)计算高密度点集合到高密度中心的距离，选择满足$\max(d(D_i, D_{\text{center}}))$的数据对象$D_i$作为第一个初始聚类中心$C_1$加入集合$C$中；

c) 选择满足$\max(d(D_j, D_{\text{center}}) \times d(D_j, C_1))$的数据对象$D_j$作为第二个初始聚类中心$C_2$加入集合$C$中；

d) 重复步骤c)，直到集合C中的元素个数等于k，即$|C| = k$；

2.2.3 更新簇中心

a) 根据式(1)计算样本集X中各数据对象与C中各中心点的距离，并按最小距离将各对象划分至最近的簇中；

b) 根据式(6),(7)得到簇内距离和矩阵S_Cluster；

c) 根据式(8)从S_Cluster中查询簇内样本距离之和最小的数据对象，并将其作为一个新的簇中心存入集合C'中；

d) 重复步骤b,c，更新各簇的中心，直到$|C'| = k$，再用C'取代C；

2.2.4 划分数据

a) 将样本集X中的数据对象划分到与其距离最近的簇中，更新簇中心集合C；

b) 计算聚类误差平方和，判断是否收敛，如果收敛，则进入“聚类评价”，否则转到“更新簇中心”；

2.2.5 聚类评价

a) 根据式(12)计算本次聚类评价指标CEI；

b) 令$k = k+1$，转到“选择初始聚类中心的步骤d)”，直至$k = k_{\max}$；

c) 根据式(13)将CEI取最大值时的k值作为最优聚类数.

2.3 DOCN 的应用流程

2.3.1 算法的基本应用

DOCN 在入侵检测系统中的应用流程如图 1，在训练阶段，通过对多个聚类结果的 *CEI* 进行对比，输出最优聚类划分(含最优聚类数和各簇中心)；在测试阶段，依次将测试集中的每条样本与各簇中心按照距离最小原则划分.

2.3.2 入侵检测模型的改进

在实际应用中，入侵检测系统可能遇到未知类型的数据，若直接按距离划分，虽然能完成聚类，但易出现偏差，导致漏检，因此，为增强算法对未知类型数据的检测，本文提出了增量式聚类更新方案，对于测试数据 T_i，假设与其距离最近的簇中心是 C_j，若 T_i 与 C_j 的距离小于等于任意两个簇中心距离的最小值，则将 T_i 划分至 C_j 所在簇，即 $T_i \in \text{Cluster}_j$，并调整该簇的聚类中心；否则，令 $k' = k+1$，跳转至聚类环节重新聚类，为避免重复运算，这里只需比较指标 $CEI(k')$ 和 $CEI(k)$ 的大小，若 $CEI(k') > CEI(k)$，则更新聚类结果，k' 为最优聚类数，否则保持原聚类结果不变.

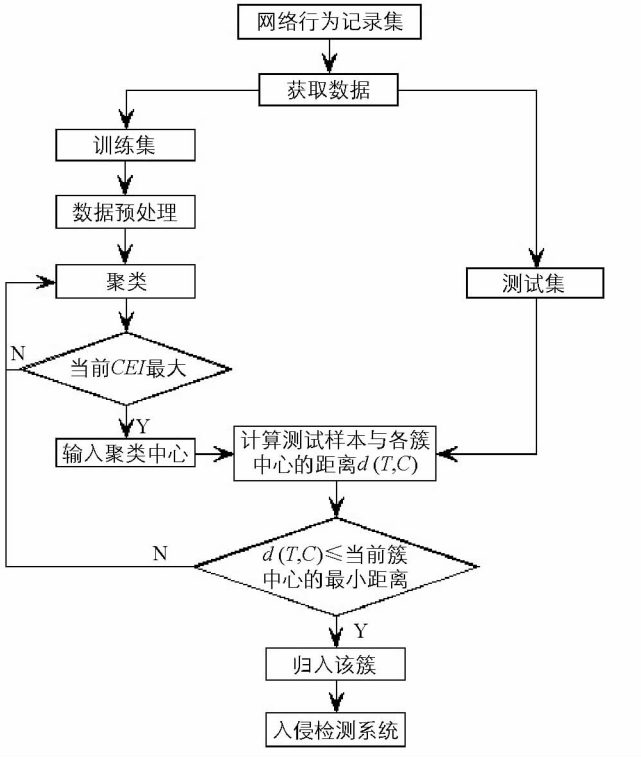

图 1　DOCN 在入侵检测系统中的应用流程

3 实验仿真与分析

为了检验 DOCN 对检测率、漏报率、误报率、分类正确率的影响，同时也为了对 DOCN 和其他 3 种算法的聚类质量进行横向比较，实验中将 K-means、文献[8]和文献[9]算法与 *CEI* 相结合，将改进后的 3 种算法依此记为“K-means+”，“文献[8]+”，“文献[9]+”.

3.1 实验环境和数据

本文实验环境为 Intel Core i5-2320 3.0 GHz，8 G 内存，1T 硬盘，Win7 操作系统，实验平台为 Matlab 2011b.

实验由两部分构成，首先，采用表 1 中的 UCI 数据集分别对改进的聚类算法和 *CEI* 指标的有效性进行了验证；其次，将改进后的入侵检测模型应用到对 KDD CUP99 网络行为记录集的异常检测中，完成聚类与评价. 实验选取 KDD CUP99 训练集中的 17 530 条记录作为训练数据，从 corrected 数据集中随机抽取两组共计 12 620 条数据作为测试集用于检验算法的性能. 为提高算法的整体运行效率，在数据预处理方面，首先使用独热编码完成字符数据的格式转换，再通过属性简约法将数据集的 41 个特征约简为 13 个[22-24]，最后将数据集归一化处理，形成新样本集，KDD CUP99 数据描述如表 2.

表 1　UCI 数据集

数据集	样本个数	属性个数	标准聚类个数
iris	150	4	3
wine	178	13	3
heart	270	13	2
wdbc	569	30	2
balance	625	4	3
sonar	208	60	2

表 2　KDD CUP99 数据集

数据集	训练集	测试集 $T1$	测试集 $T2$
Normal	12 000	4 000	4 000
DOS	3 000	1 000	1 000
PROBE	2 000	1 000	1 000
U2R	30	10	10
R2L	500	300	300

3.2 算法评估与比较

本文算法的时间复杂度为 $O(n^2+nkt)$，与其他3种算法相比，在初始聚类过程中，DOCN先计算了样本集的密度，再对高密度点集合从分布合理性的角度进行了多次筛选，虽然计算量有所增加，但各中心具有较强代表性，大体上反映了样本集的空间结构，因此可以减少后续的迭代次数；在簇中心更新过程中，DOCN使用了簇内距离和最小算法，避免了“噪音”对聚类结果的影响，而其他3种算法易受异常维度的干扰，所用均值法求得的簇中心和实际中心的位置存在偏差的可能性较大，对比结果见图2(a)，DOCN可以进一步减少迭代次数，降低算法的整体耗时，快速逼近全局最优解.

3.2.1 算法的有效性测试

在有效性验证方面，采用聚类总耗时、Rand指数、Jaccard系数和聚类准确率对K-means算法、文献[8]、文献[9]和本文算法进行了比较，考虑到K-means算法的特殊性，将算法运行20次后所得到的平均值作为其最终结果. 从图2中的实验结果可以看出：改进的聚类算法在收敛速度、聚类准确率等多项评价指标中表现良好，可以应用于实际数据的聚类.

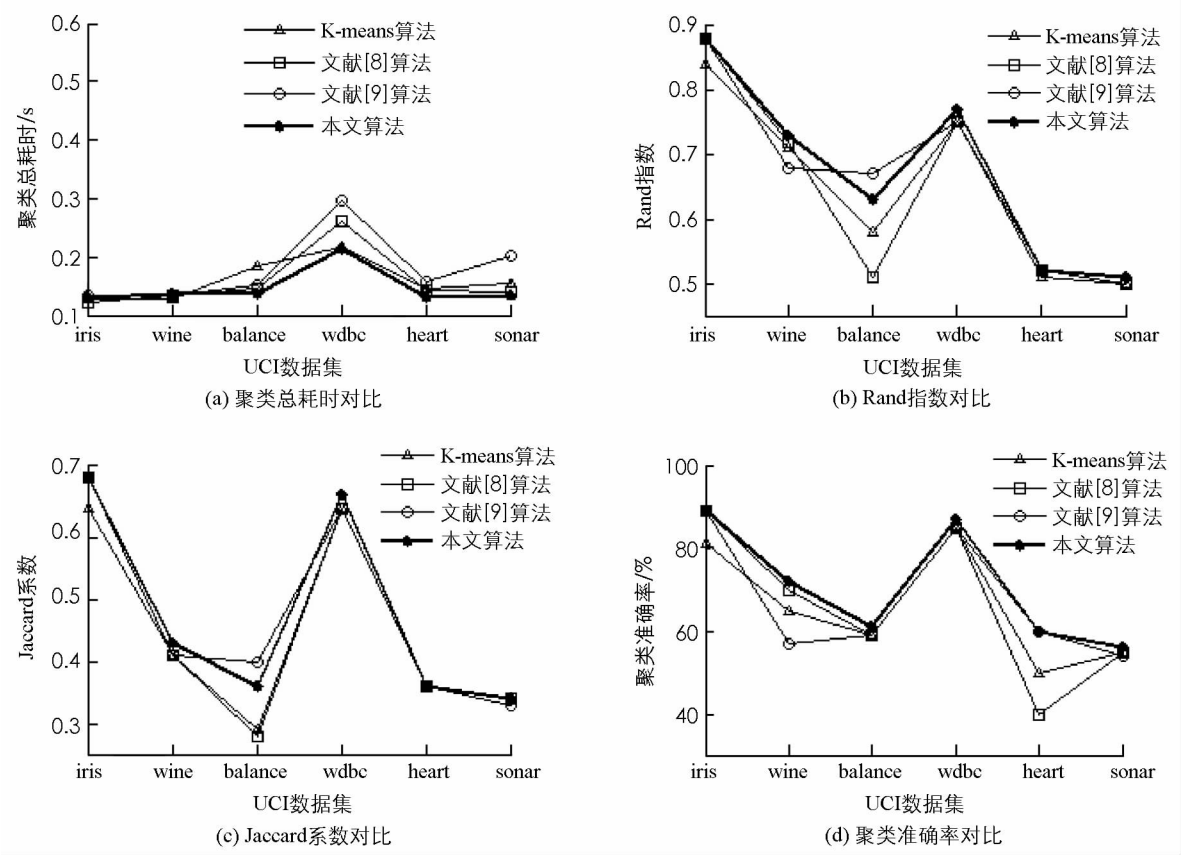

图2 外部评价指标对比结果

3.2.2 *CEI* 指标的有效性测试

对UCI数据集使用改进密度算法聚类后，得到的*CEI*评价指标如表3，可以看出*CEI*指标在6组数据上得到的聚类数目与各数据集的正确聚类数全部相符，因此该指标能够对聚类结果形成有效评价.

从上述有效性测试结果中可以看出：本文算法对聚类中心的选取合理，聚类结果理想，能够合理、准确地评价聚类质量，可以为聚类的有效评价提供重要参考依据.

3.3 算法在入侵检测中的应用

3.3.1 最优 k 值的获取

借鉴文献[12]，首先使用AP算法对样本集完成“粗聚类”，得到聚类数参考范围，在具体应用中，将参数 p 先后设置为相似度的中位数及中位数的一半，执行算法后得到训练集的聚类数范围是[15，29]，接着使用DOCN完成二次聚类和聚类评价，从图3可以看出，随着 k 值的不断增大，*CEI*逐渐增加，当 $25\leqslant k\leqslant 28$ 时，*CEI*趋于平稳，当 $k=28$ 时，*CEI*达到峰值，此后随着 k 值的再次增大，*CEI*缓慢下降.

表 3　*CEI* 指标测试结果

K 值	iris	wine	balance	wdbc	heart	sonar
2	325.470 2	313.637 1	33.112 7	2 049.055 4	186.397 6	61.384 3
3	349.206 7	360.832 3	23.630 2	953.553 9	64.036 1	19.798 6
4	309.593 0	349.685 8	35.238 2	959.008 3	132.046 5	4.635 6
5	136.810 4	326.130 2	27.182 9	965.783 3	31.159 4	14.895 4
6	115.884 3	315.706 4	29.731 8	907.722 8	56.514 8	21.920 7
7	90.244 1	336.267 6	24.014 7	877.809 3	107.215 2	30.024 7
8	62.959 4	302.228 0	26.437 7	843.285 6	63.478 0	40.517 8
9	64.262 0	349.053 6	25.215 6	845.428 7	39.589 7	40.324 6
10	61.548 8	339.096 1	22.324 7	820.144 5	25.814 1	31.308 8
11	56.615 8	347.426 5	21.920 7	814.572 4	50.965 8	28.364 8
12	51.657 8	329.887 4	17.154 6	800.452 2	48.207 4	17.847 7

3.3.2　*CEI* 对入侵检测指标的影响

在计算各项入侵检测指标环节中，将图 3 中 *CEI* 缓慢上升至峰值，再从峰值缓慢下降的这一阶段所对应的多个连续 k 值定义为最优聚类数范围，并对该范围内的各入侵检测指标进行对比.

图 4 为两个测试集的各项检测指标结果，可以看出，当聚类数在[23，26]范围内时，检测率达到峰值，分别是 93.25%，91.33%；当聚类数为 26 时，漏报率同时达到最小，分别是 2.84%，3.29%；当聚类数在[28，29]范围内时，最小误报率分别是 3.76%，4.08%；当聚类数在[28，29]范围内时，正确分类率达到最大，分别是 94.81%，93.54%.

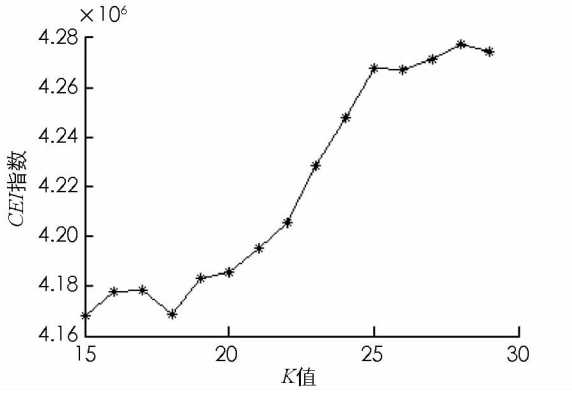

图 3　训练集的 *CEI*-*K* 的关系图

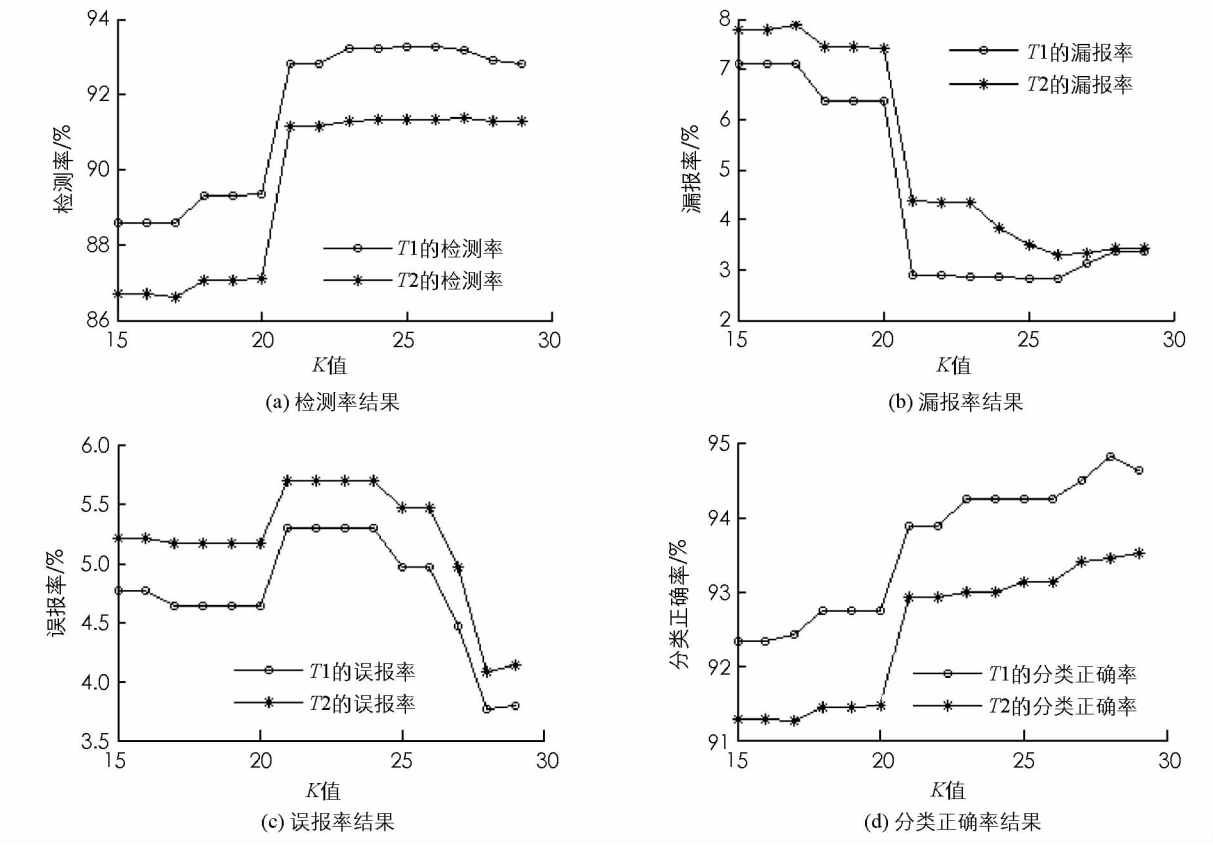

图 4　入侵检测指标结果

从图3可以看出，*CEI*越大，K值和入侵检测指标越优. 为进一步说明*CEI*与各入侵检测指标的关系，本文将训练集的*CEI*取最大值时的最优聚类数k_{opt}与测试集的各分项检测指标最优时所对应的聚类数k_{test}进行了对比，考虑到训练集与测试集都可能含有彼此未出现过的入侵数据，因此将表达式$k_{opt}=k_{test}$和$1\leqslant|k_{opt}-k_{test}|\leqslant2$分别视为二者一致和相似的条件，$k_{opt}$与$k_{test}$的分布情况如表4，经统计，$k_{opt}$与$k_{test}$的一致率为37.5%，相似率达50%.

表4 各指标最优时的K值

各指标最优条件	文献[8]+		文献[9]+		CONA算法	
	T1	T2	T1	T2	T1	T2
*CEI*最大	25	25	26	26	28	28
检测率最大	25	25	27	25	27	27
正确分类率最大	27	25	26	27	28	28
漏报率最小	26	27	17	16	26	27
误报率最小	27	16	26	27	28	28

综上所述，本文算法能够对网络行为记录集提供有效聚类，可以为入侵检测中的异常网络行为判断提供客观评价，可以为无先验知识样本集的有效聚类提供重要参考依据.

4 结 语

提出了一种基于密度和最优聚类数的改进算法，解决了无先验知识样本在聚类过程中的参数预设、聚类质量评价等问题，完成了入侵检测模型的改进. 通过引入*CEI*，给出样本的最优聚类数目，进而实现全局寻优，避免了对先验知识的依赖，为网络异常行为的识别提供一条新的研究途径. 实验表明本文算法可以在保持较低漏报率的同时，有效提高入侵检测率和分类正确率，但在实验中也发现，随着入侵检测率的提升，误报率有小幅增加的迹象. 下一步的工作将深入研究如何在提升检测率的同时保持或降低当前误报率，以及研究如何使用真实的网络数据进一步修正分类器的聚类结果，使得入侵检测模型具有更好的普适性.

参考文献：

[1] 文 华，王斐玉. 利用SSO加速最佳路径森林聚类的网络入侵检测[J]. 西南师范大学学报(自然科学版)，2017，42(5)：34－40.

[2] MACQUEEN J. Some Methods for Classification and Analysis of Mulitivariate Observations[C]//Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability. Berkeley：University of California Press，1967：281－297.

[3] PARK H S，JUN C H. A Simple and Fast Algorithm for K-medoids Clustering[J]. Expert Systems with Applications，2009，36(2)：3336－3341.

[4] LIU Y，HOU T，LIU F. Improving Fuzzy c-Means Method for Unbalanced Dataset[J]. Electronics Letters，2015，51(23)：1880－1882.

[5] 孟 静，吴锡生. 自动确定聚类数算法在网络入侵检测中的应用[J]. 计算机仿真，2013，30(10)：302－307.

[6] BREUNIG M M，KREGEL H P，NG R T，et al. LOF：Identifying Densitybased Local Outliers[J]. ACM Sigmod Record，2000，29(2)：93－104.

[7] 蒋盛益，徐雨明，陈溪辉. 异常挖掘研究综述[J]. 衡阳师范学院学报(自然科学)，2004(3)：63－66.

[8] 翟东海，鱼 江，高 飞，等. 最大距离法选取初始簇中心的K－means文本聚类算法的研究[J]. 计算机应用研究，2014，31(3)：713－715,719.

[9] 熊忠阳，陈若田，张玉芳. 一种有效的聚类中心初始化方法[J]. 计算机应用研究，2011，28(11)：4188－4190.

[10] 唐 丹，张正军. 近邻传播聚类算法的优化[J]. 计算机应用，2017，37(s1)：258－261.

[11] 倪志伟，荆婷婷，倪丽萍. 一种近邻传播的层次优化算法[J]. 计算机科学，2015，42(3)：195－200.

[12] 周世兵. 聚类分析中的最佳聚类数确定方法研究及应用[D]. 无锡：江南大学，2011.

[13] 冯柳伟，常冬霞，邓 勇，等. 最近最远得分的聚类性能评价指标[J]. 智能系统学报，2017，12(1)：67－74.

[14] 谢娟英，周　颖. 一种新聚类评价指标 [J]. 陕西师范大学学报(自然科学版)，2015，43(6)：1－8.
[15] KAPP A V，TIBSHIRANI R. Are Clusters Found in One Dataset Present in Another Dataset? [J]. Biostatistics，2007，8(1)：9－31.
[16] DAVIES D L，BOULDIN D W. A Cluster Separation Measure [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，1979，2(2)：224－227.
[17] ROUSSEEUW P J. Silhouettes：a Graphical Aid to the Interpretation and Validation of Cluster Analysis [J]. Journal of Computational and Applied Mathematics，1987，20(20)：53－65.
[18] XIE X L，BENI G. A Validity Measure for Fuzzy Clustering [J]. IEEE Transactions on Pattern Analysisand Machine Intelligence，1991，13(8)：841－847.
[19] KWON S H. Cluster Validity Index for Fuzzy Clustering [J]. Electronics Letters，1999，34(22)：2176－2177.
[20] 任　敏. 自适应模糊聚类算法及其在入侵检测中的应用研究 [D]. 济南：山东师范大学，2017.
[21] 王开军，张军英，李　丹，等. 自适应仿射传播聚类 [J]. 自动化学报，2007(12)：1242－1246.
[22] 解男男. 机器学习方法在入侵检测中的应用研究 [D]. 吉林：吉林大学，2015.
[23] 吴建胜，张文鹏，马　垣. 数据集的数据分析研究 [J]. 计算机应用与软件，2014，31(11)：321－325.
[24] 李　响. 基于经验模态分解的局域网络入侵检测算法 [J]. 西南师范大学学报(自然科学版)，2016，41(8)：132－137.

Intrusion Detection Method Based on Density and Optimal Clustering Number

ZOU Chen-song[1]，　YANG　Yu[2]

1. Department of Electrical Engineering；2. Department of Mechanical Engineering，Guangdong Songshan Polytechnic College，Shaoguan Guangdong 512126，China

Abstract：According to the problems of clustering algorithm in the application of intrusion detection，such as parameter presupposition，clustering effectiveness evaluation and unknown attack type detection，an improved algorithm based on density and optimal clustering number has been proposed. And according to the distribution of the samples，the initial clustering center has been determined heuristically，a new internal evaluation index has been proposed from the point of view of the geometric structure of the samples，and the optimal clustering number has been determined. On this basis，an incremental intrusion detection model has been designed to realize the dynamic adjustment of the clustering center and the number of clusters. Experimental results show that compared with K-means and other two improved clustering algorithms，the new algorithm has faster convergence speed and higher clustering accuracy，and can effectively cluster unknown network behaviors，and has better intrusion detection effect.

Key words：clustering algorithm；optimal clustering number；intrusion detection；effectiveness evaluation；density clustering

责任编辑　周仁惠