DOI:10.13718/j.cnki.xsxb.2017.05.006

利用 SSO 加速最佳路径森林聚类的网络入侵检测①

文 华, 王斐玉

新疆交通职业技术学院, 乌鲁木齐 831401

摘要: 针对网络入侵检测系统中的一般聚类算法速度较慢和精度较低的问题, 提出了一种基于简化群优化的最优路径森林聚类算法(SSO-OFC). 首先, 将数据集解析为图, 将其节点作为样本; 然后, 将每个样本连接到其给定特征空间中的 k-近邻, 图的节点由它们的概率密度函数(pdf)值加权得到; 最后, 通过样本及 k-近邻之间的距离计算得到 pdf 值. 提出的算法主要贡献是快速估计最佳 k 值, 并将最优路径森林聚类应用于网络入侵检测. 在 5 个公开的数据集上进行实验. 结果表明, SSO-OFC 的精度非常稳定, 除了 KddCup 数据集, 其他数据集上的精度都在 95%以上, 相比基于数据聚类的 SSO 和自组织映射更加稳定有效.

关 键 词: 网络入侵检测; 最优路径森林聚类; 简化群优化; 概率密度函数; 最佳 k 值

中图分类号: TP393 **文献标志码**: A **文章编号**: 1000-5471(2017)05-0034-07

计算机网络安全在现代信息社会的重要性不言而喻, 网络入侵检测[1]在安全基础设施中越来越重要. 一般来说, 检测分为异常和误用检测技术, 其中前一种方法用正常访问信息训练, 当待分析的新样本到来时, 系统尝试以正常访问模式匹配它, 若不匹配, 则将该样本识别为攻击[2]. 基于误用的技术由入侵(攻击)样本训练, 将任意不匹配该模式的样本分类为正常访问[3].

已有许多学者针对网络入侵检测进行了深入研究[5-7]. 文献[8]提出了一种数据聚类算法, 解析数据集为图, 其节点为样本, 且每个节点连接到给定特征空间中对应的 k-最近邻, 图的节点由它们的概率密度函数(probability density function, pdf)加权, 该方法就是著名的 OFC. 但是, OFC 中聚类效率依赖于 pdf 的估计, 且原始版本采用穷举搜索在给定区间$[k_{\min}, k_{\max}]$寻找最佳 k 值, 一般采用自然启发优化代价更小.

针对现有网络入侵检测系统中的一般聚类算法速度较慢和精度较低的问题, 提出了一种利用简化群优化加速最优路径森林聚类算法(optimum-path forest clustering using simplified swarm optimization , SSO-OFC).

1 提出算法

图 1 表示提出的算法的框架, 利用简化群优化方法替代最优路径森林聚类中的穷举搜索, 从而加速最佳 k 的计算, 通过聚类识别出攻击类型.

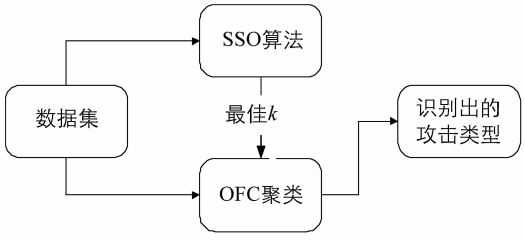


图 1 提出的算法

1.1 最优路径森林聚类

基于模式分类器设计的最优路径森林(optimum path forest, OPF)[14]是一种基于图形的方法, 利用给定特征空间中数据样本之间的连接关系, 解析训练集为图,

① 收稿日期: 2015-12-28

作者简介: 文 华(1980-), 女, 四川遂宁人, 讲师, 硕士, 主要从事计算机网络、web 安全研究.

满足给定邻接关系，其节点是样本，弧连接样本对. 在这种情况下，原型的选择、邻接关系和连接函数构成不同的分类器，本文改进了 OPF 算法[8].

令 Z 为数据集，使得每一个样品 $s \in Z$ 存在一个特征向量 $\vec{v}(s)$，令 $d(s, t)$ 为特征空间内 s 和 t 之间的距离，例如，$d(s, t) = \| \vec{v}(t) - \vec{v}(s) \|$ 即 $\vec{v}(t)$ 和 $\vec{v}(s)$ 之间的欧几里德距离. 图 (Z, A_k) 可以定义为特征空间中连接 k-最近邻的弧 $(s, t) \in A$，弧由 $d(s, t)$ 加权，节点 $s \in Z$ 由概率密度值 $\rho(s)$ 加权：

$$\rho(s) = \frac{1}{\sqrt{2\pi\sigma^2} \mid A_k(s) \mid} \sum_{\forall t \in A_k(s)} \exp\left(\frac{-d^2(s, t)}{2\sigma^2}\right) \tag{1}$$

其中：$\mid A_k(s) \mid = k$，$\sigma = \frac{d_f}{3}$，d_f 是 (Z, A_k) 中最大弧权重，这个参数选择考虑了所有用于密度计算的相邻节点，因为高斯函数覆盖 $d(s, t) \in [0, 3\sigma]$ 范围内大多数样本.

传统方法由 Parzen 窗来估计概率密度函数，式(1) 提供了基于各向同性高斯核的 Parzen 窗估计，如果 $d(s, t) \leqslant d_f$，则通过 $(s, t) \in A_k$ 定义弧. 然而，这个选择在规模和样本度中存在很多问题，针对此问题的解决方案引出依赖于特征空间区域自适应选择 d_f. 通过考虑第 k 个最近邻，该方法处理各种样本度，并降低规模问题，寻找最佳 k 值，也就是 $[k_{\min}, k_{\max}]$ 内的 $k*$，$1 \leqslant k_{\min} < k_{\max} \leqslant \mid Z \mid$.

文献[8]提出的寻找 $k*$ 的方法考虑所有聚类结果当中的最小图切割，$k \in [1, k_{\max}](k_{\min} = 1)$，归一化的度量 $GC(A_k, L, d)$：

$$GC(A_k, L, d) = \sum_{i=1}^{c} \frac{W'_i}{W_i + W'_i} \tag{2}$$

$$W_i = \sum_{\forall (s, t) \in A_k \mid L(s) = L(t) = i} \frac{1}{d(s, t)} \tag{3}$$

$$W'_i = \sum_{\forall (s, t) \in A_k \mid L(s) = i,\ L(t) \neq i} \frac{1}{d(s, t)} \tag{4}$$

其中：$L(t)$ 是样本 t 的标记，W'_i 使用聚类 i 和其他聚类之间的所有弧权重，W_i 使用聚类 $i = 1, 2, \cdots, c$ 内所有弧权重.

1.2　利用 SSO 寻找最优 k 值

下面利用简化群优化算法替代穷举搜索来加速最佳 k 的计算，即 $k*$. 在最佳路径森林聚类中，对于给定数据集 Z、距离函数 d 和值 k，邻接关系 A_k 由每个样本 $s \in Z$ 的 k 个最近邻定义. 算法 1 计算从 A_k 到 d 的概率密度函数 ρ(式(1))，在标记映射 L 中输出聚类结果. 聚类结果的质量由 $GC(A_k, L, d)$ 评估(式(2))，针对 L 中的解输出一个归一化的图割度量. 因此，将 $SSO\_OFC(Z, d, k)$ 视为一个函数，通过选择 k 来组合 Z 中的样本，并返回归一化的图割度量 $GC(A_k, L, d)$，期望找到最小化归一化图割度量的 $1 \leqslant k* \leqslant k_{\max}$.

由于优化搜索空间为一维，对于每个代理 i，$\vec{x}_i = x_i^1 = k_i$ 且 $\Phi(k_i) = GC(A_k, L, d)$，存储这些值在数组 $K[i]$，$i = 1, 2, \cdots, N$ 中，$\Phi[k]$，$k = 1, 2, \cdots, k_{\max}$. 辅助数组 $O[k]$ 标记给定 $1 \leqslant k \leqslant k_{\max}$ 的发生，避免 $SSO\_OFC(Z, d, k)$ 的重复计算. 定义 P 为给定简化群优化方法的参数集，其包括该方法的一般参数和针对每个代理的特定参数，$F(P, i)$ 是更新 i 的特定参数的函数，返回新值 k_i. 使用这些定义的 k 的优化过程如算法 1 所示.

由于 $F(P, i)$ 包括简单局部迭代，算法 1 的主要计算成本由第 9 行中函数 SSO-OFC 的执行时间 $\Theta(\mid A_{K[i]} \mid + \mid Z \mid \log(\mid Z \mid))$ 决定，对于 $1 \leqslant K[i] \leqslant k_{\max} \leqslant \mid Z \mid$，弧数 $\mid A_{K[i]} \mid = K[i] \mid Z \mid$(例如，最差情况下 $K[i] = 0.1 \mid Z \mid$). 如果 $k_{\max} \leqslant N$，则最差情况可能出现在 $SSO\_OFC(Z, d, k)$ 的 $k_{\max}$ 计算，$k = 1, 2, \cdots, k_{\max}$. 但是这仅会在主循环中出现一次，这使最差时间复杂度为 $\Theta\left(NT + k_{\max} \mid Z \mid \left(\frac{k_{\max}}{2} + \log(\mid Z \mid)\right)\right)$. 穷举搜索的时间复杂度为 $\Theta\left(k_{\max} \mid Z \mid \left(\frac{k_{\max}}{2} + \log(\mid Z \mid)\right)\right)$. 但是实践中最差情况很少出现，较少的 SSO-OFC 计算即可找到最佳 $k*$ 值.

算法1 k 的优化

输入：数据集 Z、迭代数 T、代理数 N、最大边界 k_{max}、距离函数 d、函数 SSO-OFC 和简化群优化算法的参数集合 P

输出：最佳值 $k*$

辅助：大小为 N 的数组 $K(i)$、大小为 k_{max} 的数组 $\Phi[k]$ 和 $O[k]$，变量 $\Phi*$

1. For$i=1,2,\cdots,N$, do
2. $K(i)\leftarrow\sim U(1,k_{max})$
3. $\Phi*\leftarrow+\infty$, $k*\leftarrow 0$
4. For$k=1,2,\cdots,k_{max}$, do
5. $O[k]\leftarrow$ false
6. For$t=1,2,\cdots,T$, do
7. For$i=1,2,\cdots,N$, do
8. If $O[K(i)]=$ false, then
9. $\Phi[K(i)]\leftarrow SSO\_OFC(Z,d,K[i])$ 且 $O[K(i)]\leftarrow$ true
10. If $\Phi*>\Phi[K[i]]$, then
11. $\Phi*\leftarrow\Phi[K[i]]$ 且 $k*\leftarrow K[i]$
12. 更新 $K[i]$,$\Phi*$,$\Phi[i]$ 和 P 中 $k*$
13. For$i=1,2,\cdots,N$, do
14. $K[i]\leftarrow F(P,i)$

图2表示真实场景中提出的算法，假设系统管理员面临一个计算机网络中的不寻常场景，如假阳性(分类攻击为正常访问)数增加，则可以手动分类这类样本为某类攻击，然后添加该信息到数据集，使用 SSO-OFC 进行离线学习. 现在由于聚类数可能不同(新攻击种类的情况下，可能会形成新聚类)，必须再一次学习 $k*$ 值，则新分类器会再返回，并假设对这类攻击更具鲁棒性.

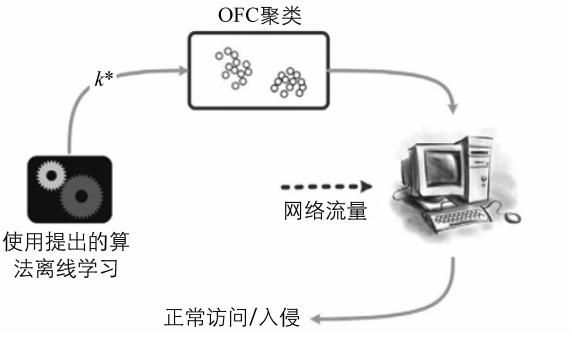


图2 真实场景中提出的算法

2 实 验

采用了5个公开数据集(IDS_Bag[15], KddCup[16], Netflow[5], ISCX[4], NSL-Kdd[6])来评估算法的有效性.

2.1 最佳 $k*$ 值估计

为了评估通过 SSO 寻找最佳 $k*$ 值的性能，首先评估穷举搜索算法的行为. 图3展示了穷举搜索算法在 KddCup,NSL-Kdd,Netflow 和 ISCX 数据集上的曲线. 穷举搜索在范围[1, 100]内执行，步长为1. 可以观察出，KddCup,NSL-Kdd 的最小图割度量分别为 3.900 435,0.000 013. 关于 Netflow 数据集，针对 $k\geqslant 70$，图割为 0.0，即该数据集相对于该度量非常好. 相对于 ISCX 数据集，在 $k=80$ 获得的最小图割是 0.000 028，表明 SSO-OFC 可在该数据集上获得良好结果.

图4－6表示使用本文提出的简化群优化估计最佳 $k*$ 值执行20次迭代的结果. 在较高迭代次数时，代理可能会减慢算法，因此，需要权衡“迭代×代理”在提供计算负载和搜索能力之间的适当折中. 为每个代理数执行算法5次，平均它们的最小图割值和执行时间，以及它们各自的标准差. 由图6可以看出，关于 ISCX 数据集，20次迭代似乎适合考虑5,10和15个代理的 FFA 和 HS 算法.

由图4可以看出，KddCup 数据集上最慢的技术 BA 比穷举搜索快1.7倍. 由图5可以看出，NSL-Kdd 数据集上最慢的技术(BA)比穷举搜索快1.65倍. 最好的情况下，即5个代理的情况，BA 是最慢的技术，但在两个数据集上，其速度约比穷举搜索快2倍.

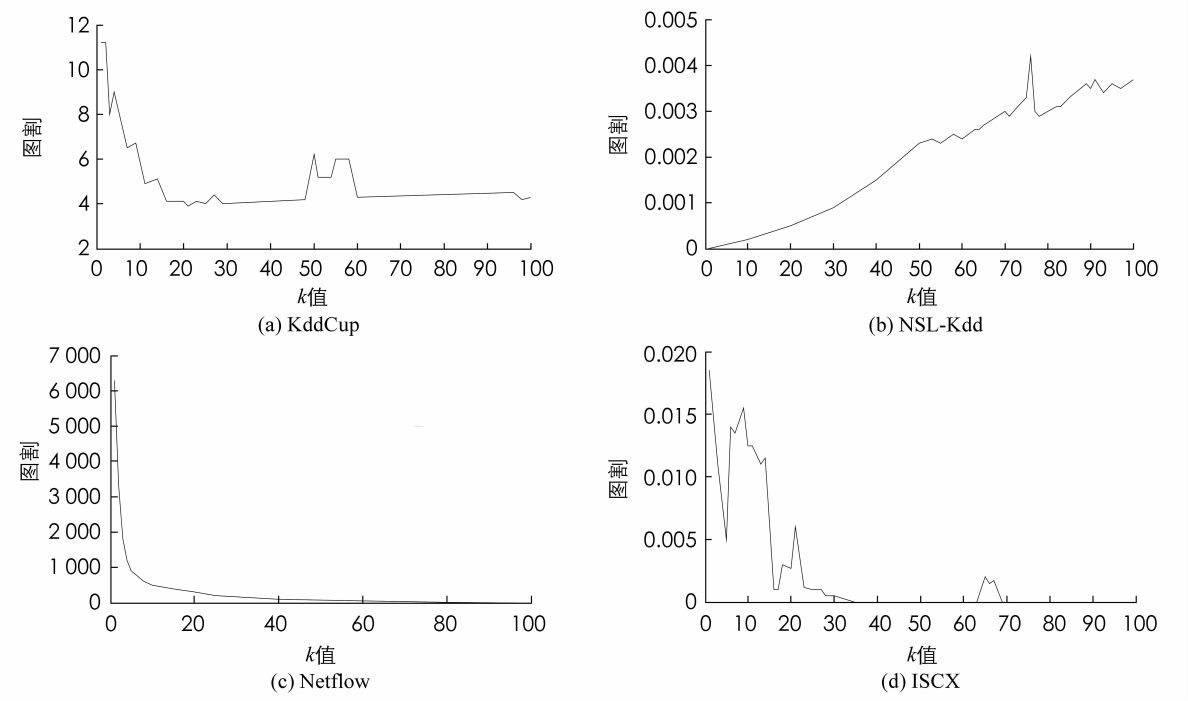


图 3　穷举搜索算法在各个数据集上的结果

尽管 HS 算法在各种代理数和迭代数下是最快技术，但是它不能在任意数据集上实现合适的图割值(见图 4(a)，图 5(a)和图 6(a))，这可能是由于小迭代数和离散搜索空间域所致. 此外，SSO 在寻找最小图割度量的任务中优于其他几种算法，且需要最少的平均执行时间，表明 SSO 在估计最佳 $k*$ 值方面具有明显优势. SSO 的很大优势在于它的更新策略，它增加了随机惯性权重，使粒子可以反向搜索，不容易陷入局部最优，提高了收敛速度，因此获得了最好的图割值. 其他几种方法都容易陷入局部最优值且速度较慢.

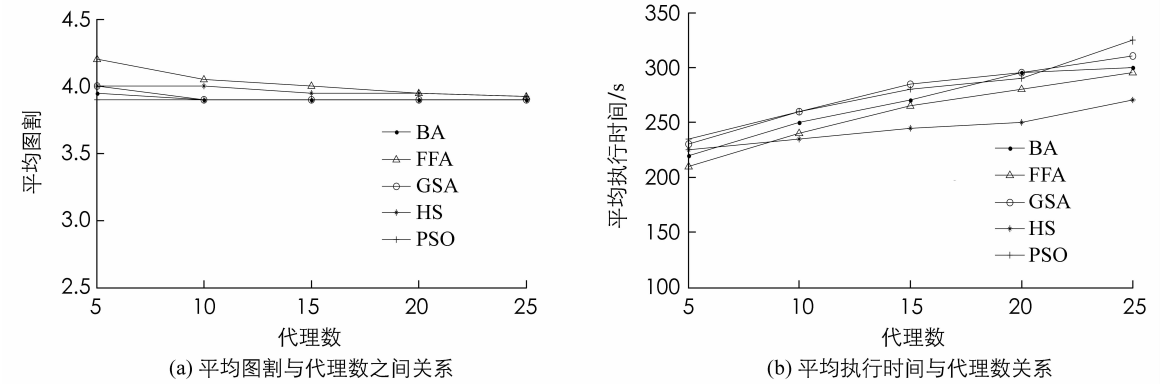


图 4　KddCup 数据集上的实验结果

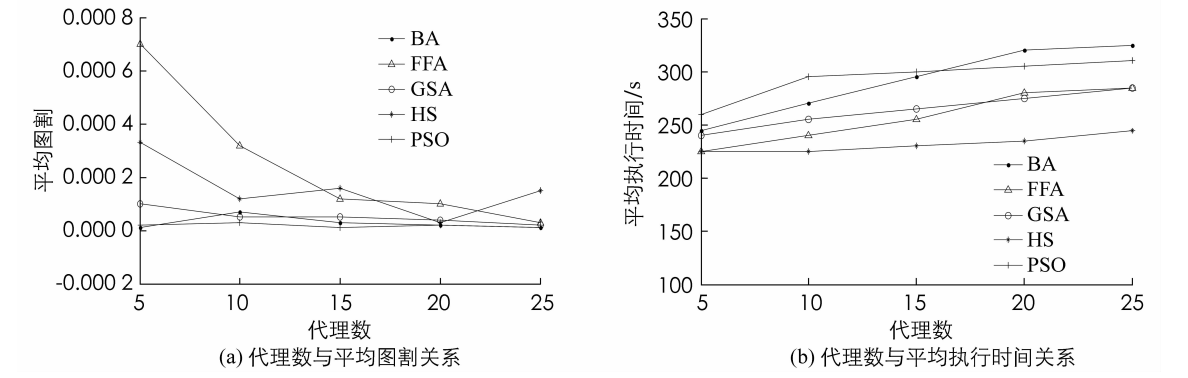
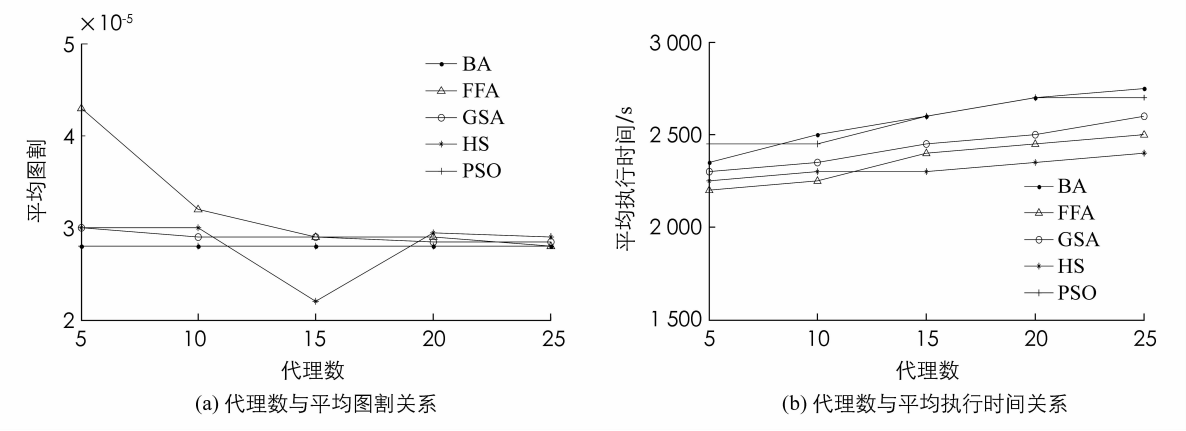


图 5 NSL-Kdd 数据的实验结果



(a) 代理数与平均图割关系 (b) 代理数与平均执行时间关系

图 6 ISCX 数据集上的实验结果

2.2 与其他聚类算法进行比较

将 SSO-OFC 与基于聚类的 SOM,SSO 进行比较，对 SOM 和 SSO 使用本文的实现方法. 根据经验选择元启发式参数，在下面各种情景下评估了优化技术的性能：不同迭代次数(1～20)，不同代理数(5～20)，步长为 5. 表 1 为在 ISCX 上运行 20 次的分类精度，从表 1 中可看出，本文算法分类精度最高. 通常分类精度取决于算法对于数据集特征的分类性能，本文算法的主要环节是最佳 k 值的计算，这也是对最优路径森林算法的一种改进. 之所以选择 SSO 用于 k 值计算，因为 SSO 是一种优化的粒子群算法，可以动态选择特征因子，在上一小节中最佳 k 值估计已经得到了证明. SSO-OFC 在特征空间自适应选择最大权重，通过考虑最近邻域，降低了规模问题，变相提高了分类精度.

表 2 为针对 IDS_Bag,KddCup,NSL-Kdd,Netflow 和 ISCX 数据集的实验结果. 总体而言，提出的算法在大部分数据集上获得了最高的检测精度，只是在 IDS-Tuesday 数据集中低于 SOM.

表 1 在 ISCX 运行 20 次的分类精度

运行次数/次	SSO-OFC 验证平均/%	SSO 验证平均/%	SOM 验证平均/%
1	93.4	89.9	88.8
4	93.5	90.0	86.7
8	93.4	89.1	89.2
9	93.5	90.0	89.0
10	93.4	87.5	88.9
12	93.3	89.5	89.0
13	93.2	89.9	89.2
14	93.4	90.0	89.2
15	93.3	90.0	89.0
16	93.0	89.6	89.2
18	93.5	89.7	86.2
20	93.4	90.0	88.9
最小	92.5	87.5	86.1
最大	93.5	90.0	89.2
平均	93.3	89.6	88.5
标准差	0.23	0.72	1.07

表 2　所有数据集的检测精度

数据集	SSO-OFC	SSO	SOM
IDS-Monday	1.000 0	1.000 0	1.000 0
IDS-Tuesday	0.969 0	0.995 5	0.995 5
IDS-Thursday	0.997 3	0.980 5	0.997 3
IDS-Friday	0.988 0	0.942 3	0.888 4
ISCX	0.963 7	0.963 45	0.963 4
KddCup	0.716 6	0.600 71	0.600 1
NSL-Kdd	0.998 8	0.534 63	0.540 3
Netflow	0.957 7	0.759 45	0.214 5

3　结　　语

本文引入了计算机安全领域中的最优路径森林聚类，用于检测恶意或异常行为，首先在 5 个公开数据集上进行实验，在计算机网络入侵检测环境下评估了 SSO-OFC 与 SOM 网络以及 k -均值的鲁棒性，SSO-OFC 在一半数据集中优于所有技术. 然后，本文评估了多种自然启发优化技术，寻找用于 OFC 计算的参数 k＊，所有技术均比穷举搜索快，而 SSO 取得了最好的性能.

未来会将提出的算法应用于其他数据集，并考虑与其他的优化算法相结合，如模拟退火算法[17]、遗传算法[18]等，通过实验进一步改善算法的效率和检测精度.

参考文献：

[1] 彭茂玲，陈善雄，余光琳. 一种基于压缩感知的入侵检测方法 [J]. 西南大学学报(自然科学版)，2014，36(2)：186－192.

[2] HUANG M，CAI Z G. Design and Realization of Network Instruction Platform Based on Role-Playing [C]// Computer Science & Education (ICCSE)，2014 9th International Conference. New York：IEEE Press，2014：953－956.

[3] 田志宏，王佰玲，张伟哲，等. 基于上下文验证的网络入侵检测模型 [J]. 计算机研究与发展，2013，50(3)：498－508.

[4] 陈岳兵. 面向入侵检测的人工免疫系统研究 [D]. 长沙：国防科学技术大学，2011.

[5] Kozak J，Boryczka U. Multiple Boosting in the Ant Colony Decision Forest meta-classifier [J]. Knowledge-Based Systems，2015，25(2)：141－151.

[6] CHABAA S，ZEROUAL A，ANTARI J. Identification and Prediction of Internet Traffic Using Artificial Neural Networks [J]. Journal of Intelligent Learning Systems & Applications，2010，17(3)：147－155.

[7] YANG H，XIE X，WANG R. SOM-GA-SVM Detection Based Spectrum Sensing in Cognitive Radio [C]//Wireless Communications，Networking and Mobile Computing (WiCOM)，2012 8th International Conference. New York：IEEE Press，2012：1－7.

[8] ROCHA L M，CAPPABIANCO F A M，FALCAO A X. Data Clustering as an Optimum-Path Forest Problem with Applications in Image Analysis [J]. International Journal of Imaging Systems and Technology，2009，19(2)：50－68.

[9] YANG XIN-SHE，HE XINGSHI，Bat Algorithm：Literature Review and Applications [J]. Int J Bio-Inspired Computer，2013，5(3)：141－149.

[10] 黄浦博，尉　宇. 改进的萤火虫群优化算法及其非线性盲源分离 [J]. 电信科学，2015，31(9)：97－102.

[11] XU B C，ZHANG Y Y. An Improved Gravitational Search Algorithm for Dynamic Neural Network Identification [J]. International Journal of Automation & Computing，2014，11(4)：434－440.

[12] 李永林，叶春明，刘长平. 轮盘赌选择自适应和声搜索算法 [J]. 计算机应用研究，2014，31(6)：1665－1668.

[13] CHUNG Y Y，WAHID N. A Hybrid Network Intrusion Detection System Using Simplified Swarm Optimization (SSO) [J]. Applied Soft Computing，2012，12(9)：3014－3022.

[14] SOUZA R，RITTNER L，LOTUFO R. A Comparison Between k-Optimum Path Forest and k-Nearest Neighbors Supervised Classifiers [J]. Pattern Recognition Letters，2014，39(4)：2－10.

[15] YEH W, CHANG W, CHUNG Y Y. A New Hybrid Approach for Mining Breast Cancer Pattern Using Discrete Particle Swarm Optimization and Statistical Method [J]. Expert Systems with Applications, 2009, 36(4): 8204 - 8211.

[16] SHAH B, TRIVEDI B H. Reducing Features of KDD CUP 1999 Dataset for Anomaly Detection Using Back Propagation Neural Network [C]// Advanced Computing & Communication Technologies (ACCT), 2015 Fifth International Conference. New York: IEEE Press, 2015: 247－251.

[17] 杨建辉，吴　聪. PSO结合SA优化算法的无线传感器网络路由协议 [J]. 湘潭大学学报(自然科学版)，2015，37(4)：98－104.

[18] 杨　建，刘述木，黎远松. 一种基于改进遗传算法的WSN负载均衡聚类算法 [J]. 西南师范大学学报(自然科学版)，2015，40(10)：41－48.

On a Speed up Optimum-Path Forest Clustering Algorithm Using SSO for Network IDS

WEN　Hua，　WANG Fei-yu

Xinjiang Vocational and Technical College of Communications, Urumqi Xinjiang 831401, China

Abstract: Concerning that the general clustering algorithms in network intrusion detection systems is slow and the accuracy rate is low, an optimum-path forest clustering algorithm (OFC) using simplified swarm optimization (SSO) has been proposed. Firstly, OFC analyzes data sets in figures, with the graphic nodes being as the samples. Then, every sample is connected to its feature space given *k*-nearest neighbor (KNN chart). The graph nodes are weighted by their probability density function (pdf). Finally, the pdf values are obtained by calculating the distance between the sample and the *k*-nearest neighbor. The main contribution of the proposed algorithm is to estimate the optimal k value fast, and the SSO-OFC is used in network intrusion detection. The experimental results on five public data sets show that the accuracy performance on five databases is very stable. Except on KddCup, the accuracy on other data sets are all above 95%, much more stable than clustering based on SSO and self-organizing map.

Key words: network intrusion detection; optimum-path forest clustering; simplified swarm optimization; probability density function; optimal k value

责任编辑　张　枸