PSO-based K-means算法及其在网络入侵检测中的应用

傅 涛 孙文静

(南京审计学院计算机系 南京 210029) (江苏博智软件科技有限公司 南京 210012)

摘 要 PSO算法是一种基于群体智能的群优化和群搜索算法,效率高、收敛快。提出将其与K-means算法结合,用于网络入侵检测。实验表明,PSO-based K-means算法克服了K-means算法对初始聚类中心、孤立点和噪声敏感且易陷入局部最优解的缺点,收敛速度快,检测准确率较高。

关键词 入侵检测,PSO,K-means算法,检测准确率

中图法分类号 TP311 **文献标识码** A

PSO-based K-means Algorithm and its Application in Network Intrusion Detection

FU Tao SUN Wen-jing

(Department of Computer, Nanjing Audit University, Nanjing 210029, China)

(Jiangsu Bosh Company of Software & Technology, Nanjing 210012, China)

Abstract PSO is an algorithm based on swarm intelligence optimization and search, has high efficiency, fast convergence. In this paper, it combines with the K-means algorithm for network intrusion detection. Experiment shows that PSO-based K-means algorithm overcomes the shortcoming that the K-means algorithm is sensitive to the initial cluster centers, outliers and noise, easy to fall into local optimal solution. It's an algorithm with fast convergence and higher detection accuracy.

Keywords Intrusion detection, PSO, K-means algorithm, Detection accuracy

K-means算法是一种解决聚类问题的经典算法,简单、快速,对处理大数据集具有可伸缩性,在类与类之间区别明显时,其聚类效果较好。但是,K-means算法仅用欧式距离作为相似性度量的标准,对初始聚类中心、孤立点和噪声敏感,易陷入局部最优解。

针对这一缺点,本文提出PSO-based K-means算法,并将其用于网络入侵检测。实验表明,PSO-based K-means算法检测率远高于传统K-means算法,同时网络攻击误报率及漏报率有明显降低,这说明,本文提出的PSO-based K-means算法有良好的网络攻击检测能力。

1 PSO-based K-means算法

1.1 K-means算法

K-means算法的基本思想是:随机地选择 k 个对象,每个对象初始地代表了一个类的平均值或中心。对剩余的每个对象根据其与各个类中心的距离,将它赋给最近的类,然后重新计算每个类的平均值。不断重复该过程,直到准则函数收敛。准则函数定义为:

$$E=\sum_{i=1}^{k}\sum_{x\in C_i}|x-\overline{x_i}|^2 \tag{1}$$

式中,E 是数据集所有对象与它所在类的中心点的平方误差的总和,E 越大说明对象与聚类中心的距离越大,类内的相似性越低;E 越小说明类内的相似性越高。x 为类内的一个数据对象;$\overline{x_i}$ 是类 C_i 的聚类中心;k 是类的个数;C_i 是 k 个类中的第 i 个类。

K-means算法采用欧几里德距离作为相似性度量的标准:

$$d(i,j)=\sqrt{(x_{i1}-y_{j1})^2+(x_{i2}-y_{j2})^2+\cdots+(x_{in}-y_{jn})^2} \tag{2}$$

式中,$d(i,j)$ 是 n 维数据对象 i 和 j 的距离;$x_{i1},x_{i2},\cdots,x_{in}$ 和 $y_{j1},y_{j2},\cdots,y_{jn}$ 分别表示对象 i 和对象 j 在第 $1,2,\cdots,n$ 维的数据。

K-means算法如下:

输入:类的数目k和包含n个对象的数据集;

输出:E收敛或是聚类中心不再变化的k个聚类。

算法步骤:

(1)随机选择k个数据对象作为初始聚类中心;

(2)Repeat;

(3)对每个对象根据其与聚类中心的距离,将其划分给距离最近的类;

(4)重新计算每个类所有对象的平均值,得到每个类的新的聚类中心;

(5)Until E收敛或是聚类中心不再发生改变。

1.2 PSO-based K-means算法

PSO(Particle Swarm Optimization)算法[1]是一种基于群体智能的群优化和群搜索算法,目前被广泛应用于函数优化、

到稿日期:2013-01-25 返修日期:2013-04-11 本文受国家发改委发改办[2012]3179号下一代互联网络扫描与补丁管理系统产业化项目基金资助。

傅 涛(1980—),男,博士,助理研究员,主要研究方向为信息安全;孙文静 女,硕士,讲师,主要研究方向为计算机网络。

神经网络训练、模式分类、模糊系统控制及其他的应用领域[2]。

PSO算法模拟鸟群的捕食行为，通过鸟之间的集体协作使群体达到最优目的。PSO算法的基本思想是：每个优化问题的解称为 n 维空间的一个粒子，每个粒子都有初始位置和速度，通过适应度函数计算粒子的适应度值，用来衡量粒子距离目标位置的远近，粒子根据自身的飞行经验和粒子群的飞行经验动态调整飞行速度和位置，追逐当前在解空间中的最优粒子来搜索最好的目标位置，从而得到所优化问题的最优解[3]。

PSO算法保留了基于种群的全局搜索策略，但其采用的速度-位移模型操作简单，避免了复杂的遗传操作。它特有的记忆使其可以动态跟踪当前的搜索情况来调整搜索策略，是一种更高效的并行搜索算法[4]。

PSO算法的基本概念定义如下：

定义 1(粒子的位置和速度) 粒子群中第 i 个粒子在 n 维空间的位置表示为：$x_i=(x_{i1},x_{i2},\cdots,x_{in})$，速度表示为：$v_i=(v_{i1},v_{i2},\cdots,v_{in})$。

定义 2(粒子经历过的最好位置) 粒子群中第 i 个粒子在 n 维空间中经历过的最好位置表示为：$p_i=(p_{i1},p_{i2},\cdots,p_{in})$。

定义 3 整个粒子群经历过的最好位置表示为：$g=(g_1,g_2,\cdots,g_n)$。

定义 4 粒子更新自身速度和位置的公式为：

$$v_i=\omega * v_i+acc\_const * rand * (p_i-x_i)+acc\_const * rand * (g-x_i) \quad (3)$$

$$x_i=x_i+v_i \quad (4)$$

式中，ω 为权重系数，acc_const 为加速因子，通常取值为 2，$rand$ 取(0,1)区间的随机数。

PSO-based K-means算法描述如下：

输入：包含 n 个对象的数据集和最大迭代次数 MaxIter；

输出：优化后的数据集。

算法步骤：

(1)每个数据集对象为一个粒子，随机设定每个粒子的初始位置 x_i 和初始速度 v_i；

(2)Repeat；

(3)根据适应度函数计算每个粒子的适应度值 $f(x_i)$；

(4)比较每个粒子的适应度值与它经历过的最好位置 p_i 的适应度值，如果适应度值更大，则用当前粒子的位置和适应度值更新 p_i 和 p_i 的适应度值；

(5)比较每个粒子的适应度值与整个粒子群经历过的最好位置 g 的适应度值，如果适应度值更大，则用当前粒子的位置和适应度值更新 g 和 g 的适应度值；

(6)根据式(3)和式(4)更新每个粒子的速度和位置；

(7)Until 达到最大迭代次数 MaxIter；

(8)从得到的这 k 个初始聚类中心出发，运用 K-means 算法进行聚类。

1.3 算法实验及结果分析

本文将传统的 K-means 算法和 PSO-based K-means 算法进行了对比实验。采用的测试数据集是 UCI[46] 数据库的 Iris、Glass Identification、Wine 和 Hayes-Roth 4 组数据，测试数据集的数据分布保持原有状态，不做任何的人工处理。

实验中，为上述 4 组数据集 K-means 算法指定的领域半径 ε 分别为 1、10000、40 和 500。对于 PSO-based K-means 算法，指定最大迭代次数 MaxIter 为 20。

实验在 CPU：Inter Core2 E7300 2.66GHz，内存：2G，OS：Window XP Professional，开发环境：Eclipse3.5 的环境下进行。运用传统 K-means 算法进行 20 次实验，PSO-based K-means 算法进行 1 次实验，实验结果如表 1 所列。

表 1 K-means 算法与 PSO-based K-means 算法聚类精度比较

算法	数据集	最高聚类精度	最低聚类精度	平均聚类精度
K-means	Iris	89.33%	82.00%	87.30%
	Glass	74.77%	47.20%	64.74%
	Wine	74.16%	70.22%	71.01%
	Hayes-Roth	80.30%	77.27%	78.41%
基于 PSO	Iris	89.33%	89.33%	89.33%
	Glass	75.01%	57.01%	67.02%
	Wine	70.22%	70.22%	70.22%
	Hayes-Roth	80.30%	80.30%	80.30%

由表 1 可见，对 Iris、Glass 和 Hayes-Roth 数据集，PSO-based K-means 算法的最低聚类精度、最高聚类精度和平均聚类精度均高于 K-means 算法，这说明在大多数情况下，PSO-based K-means 算法的聚类效果优于传统 K-means 算法。

表 1 中，对于 Wine 数据集，PSO-based K-means 算法的聚类精度较传统的 K-means 算法的平均聚类精度低，只能达到传统算法的最低聚类精度。初步分析，Wine 数据集的每个数据项有 14 个属性，各个数据项的取值范围差距较大，尤其是 Proline 属性，它的取值范围是(278～1680)。对于 PSO-based K-means 算法，Proline 属性取值范围大，影响了 PSO 算法优化得到的初始聚类中心，使得全局优化效果不好，这说明改进的 K-means 算法有一定的局限性。

2 PSO-based K-means 算法在网络入侵检测中的应用

图 1 是一种常用的入侵检测系统(IDS)模型。IDS 首先从网络上捕获数据包，对数据包进行预处理；然后对待检测的数据进行误用检测，与入侵行为规则库进行匹配，能够匹配的数据为具有入侵行为的数据，对这些数据及时做出响应；最后将不能匹配的数据进行异常检测，能够与正常行为规则库匹配的数据为具有正常行为的数据，对于不能匹配的数据及时做出响应，然后做规则挖掘处理，将产生的新的入侵行为规则添加到入侵行为规则库中，将新的正常行为规则添加到正常行为规则库中。

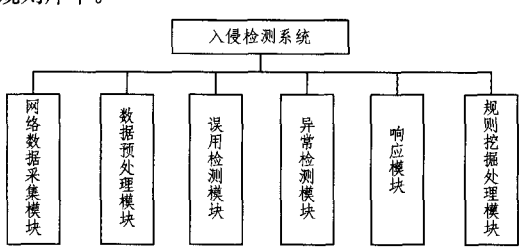

图 1 入侵检测系统总体结构

图 2 为规则挖掘处理模块的流程图。规则挖掘处理过程的步骤如下：

(1)将异常检测后剩余的数据保留为待规则挖掘的数据；

(2)运用聚类算法对这些数据进行处理，挖掘新的行为规则，将新产生的入侵行为规则添加至入侵行为规则库，将新产生的正常行为规则添加至正常行为规则库，规则挖掘过程结束。

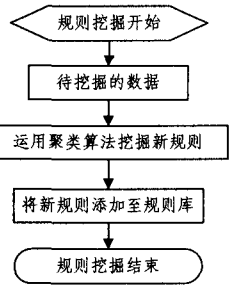

图 2　规则挖掘处理流程图

本文运用 PSO-based K-means 算法挖掘新的行为规则。

实验采用的数据集是 KDD Cup1999[5] 数据集中的"kddcup. data_10_percent",该数据集是从一个模拟的美国空军局域网上采集的 9 个星期的网络连接数据,是入侵检测领域公认的权威测试数据集。该数据集共有 494021 条记录,其中正常记录条数 97278,其余为入侵记录条数。攻击类型分为 4 类:非法企图中断或干扰主机或网络的正常运行的攻击(DoS(Denial of service))、非法扫描和探测主机和网络的攻击(Probe)、远程未授权用户非法获得本地主机的用户权限的攻击(R2L(Remote to Local))和本地非授权用户非法获取本地超级用户或管理员权限的攻击(U2R(User to Root))。每条数据有 33 个数值型属性和 8 个符号型属性,提供了 41 个特征,分为 4 类,分别是基本特征、流量特征、内容特征、主机流量特征[6]。

将 K-means 算法和 PSO-based K-means 算法进行了对比实验。K-means 算法进行 20 次实验取平均值,PSO-based K-means 算法的最大迭代次数 MaxIter 为 10。实验结果如表 2 —表 5 所列。表中,()内数值对应 PSO-based K-means 算法的检测结果。检测率、误报率和漏报率分别为:

检测率=检测出的攻击记录数/总攻击记录数×100%　(5)

误报率=误认为攻击的正常记录数/总正常记录数×100%　(6)

漏报率=未检测出的攻击记录数/总攻击记录数×100%　(7)

表 2　DoS 类型数据的入侵检测结果

样本个数	检测率	误报率	漏报率
4000	85.96%(83.33%)	23.53%(0.03%)	14.04%(16.67%)
6000	84.47%(82.22%)	18.17%(0.02%)	15.53%(17.78%)
8000	81.27%(73.33%)	11.56%(0.02%)	18.73%(26.67%)
10000	81.75%(81.33%)	15.84%(2.32%)	18.25%(18.67%)

表 3　Probe 类型数据的入侵检测结果

样本个数	检测率	误报率	漏报率
4000	91.75%(94.17%)	4.58%(0.10%)	8.25%(5.83%)
6000	90.47%(96.11%)	8.29%(0.14%)	9.53%(3.89%)
8000	95.42%(97.08%)	4.07%(0.10%)	4.58%(2.92%)
10000	96.67%(100.00%)	1.14%(0.39%)	3.33%(0.10%)

表 4　R2L 类型数据的入侵检测结果

样本个数	检测率	误报率	漏报率
4000	60.54%(94.17%)	1.89%(0.10%)	39.46%(5.83%)
6000	72.92%(94.44%)	2.65%(0.02%)	27.08%(5.56%)
8000	78.25%(96.67%)	2.50%(0.06%)	21.75%(3.33%)
10000	71.28%(97.33%)	17.48%(0.08%)	28.72%(2.67%)

表 5　U2R 类型数据的入侵检测结果

样本个数	检测率	误报率	漏报率
4000	88.33%(93.33%)	3.59%(4.36%)	11.67%(6.67%)
6000	95.53%(86.67%)	3.25%(0.03%)	4.47%(13.33%)
8000	96.27%(96.67%)	3.13%(3.26%)	3.73%(3.33%)
10000	97.70%(99.33%)	3.36%(0.14%)	2.30%(0.67%)

表 6 是两种算法入侵检测结果的对比情况,其中:

检测率提高率=(PSO-based K-means 算法检测结果−K-means 算法检测结果)/ K-means 算法检测结果

表 6　K-means 算法和 PSO-based K-means 算法的入侵检测结果对比

攻击类型	检测率提高率(%)(P_1,P_2,P_3,P_4)	误报率降低率(%)(F_1,F_2,F_3,F_4)	漏报率降低率(%)(T_1,T_2,T_3,T_4)
DoS	−4.2,−2.66, −9.77,−0.51	99.87,98.27, 99.83,85.35	−18.73,59.18 −42.39,−2.3
Probe	2.56,29.51 1.74,3.44	97.82,98.31 97.54,65.79	29.33,59.18 36.24,97.00
R2L	54.43,22.79 23.54,26.15	94.7,99.25 −3.99,99.54	85.22,79.47 84.69,90.70
U2R	12.45,−9.27 0.41,1.67	−17.66,99.08 95.83,95.82	42.84,−66.47 10.72,70.87

表 6 中,检测率提高率为(PSO-based K-means 算法检测结果−K-means 算法检测结果)/K-means 算法检测结果,而误报率降低率、漏报率降低率则为(K-means 算法检测结果−PSO-based K-means 算法检测结果)/ K-means 算法检测结果,P_i、F_i、T_i 下标 1、2、3、4 分别对应样本个数 4000、6000、8000 和 10000。由表 6 可见,除 DoS 攻击外,PSO-based K-means 算法检测率远高于传统算法,同时误报率及漏报率也有明显降低,这说明,本文提出的 PSO-based K-means 算法有良好的网络攻击检测能力。

结束语　传统 K-means 算法对初始聚类中心敏感,容易陷入局部最优解。针对这一不足,本文提出 PSO-based K-means 算法,其具有全局搜索能力,得到的聚类结果不会像传统 K-mean 算法那样陷入局部最优解。此外,它特有的记忆使其可以动态跟踪当前的搜索情况来调整搜索策略,是一种更高效的并行搜索算法。将 PSO-based K-means 算法用于网络入侵检测,可有效提高检测准确率,降低误报率与漏报率。

参 考 文 献

[1] Kennedy J, Eberhart R. Particle Swarm Optimization[C]//Proceedings of IEEE International Conference on Neural Networks. 1995, 4: 1942-1948

[2] 田东平,徐成虎. 改进的粒子群优化算法的研究和分析[J]. 计算机工程与应用,2008,44(34):56-60

[3] 赵昌. 基于粒子群优化的入侵检测规则提取方法研究[D]. 合肥:中国科学技术大学,2009:14-15

[4] 周驰,高海兵,高亮,等. 粒子群优化算法[J]. 计算机应用研究,2003,12:7-11

[5] ACM Special Interest Group on Knowledge Discovery and Data Mining. KDD Cup 1999[DB/OL]. http://www.sigkdd.org/kddcup/index.php? section=1999&method=data

[6] 谷保平. 基于聚类算法的网络入侵检测研究[D]. 广州:广州工业大学,2008:35-37