基于PSO的k-means算法及其在网络入侵检测中的应用

傅 涛 孙亚民

(南京理工大学计算机与技术学院 南京 210094)

摘 要 在传统k-means算法中,初始聚类中心随机选择,聚类结果随初始聚类中心的不同而波动,从而导致聚类结果不稳定。提出的PSO-based k-means算法使用PSO算法优化生成初始聚类中心,得到的聚类结果全局最优,不会陷入局部最优解。实验结果表明,将PSO-based k-means算法用于入侵检测系统的规则挖掘处理模块,其入侵检测率明显高于传统k-means算法,而误报率则大大低于后者。显然,PSO-based k-means算法可有效提高网络入侵检测系统的性能。

关键词 PSO-based k-means,优化聚类,入侵检测,检测率,误报率

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

FU Tao SUN Ya-min

(Dept. Computer Science and Technology, Nanjing University of Science and Technology, Nanjing 210094, China)

Abstract In the traditional k-means algorithm, the initial cluster center is chosen randomly, clustering result varies from the initial cluster center, and clustering result is unstable. The PSO-based k-means algorithm was proposed in the paper. The PSO optimization algorithm generates the initial cluster center. The clustering result is global optimal and doesn't fall into local optimal solution. Experimental results show that the intrusion detection rate is significantly higher than the traditional k-means algorithm and its false positive rate is largely lower than the latter by applying the PSO-based k-means algorithm to the rule mining module of intrusion detection system. Obviously, the PSO-based k-means algorithm can improve the performance of network intrusion detection system effectively.

Keywords PSO-based k-means, Optimization clustering, Intrusion detection, Detection rate, False positive rate

1 引言

k-means算法是一种得到最广泛使用的基于划分的聚类算法[1],其基本思想是: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,反之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,…,$n$维的数据。

k-means算法简单、快速,当类与类之间区别明显时,该算法聚类效果较好。但是k-means算法只有在类的平均值被定义的前提下才能使用,算法要求用户事先给出类的数目k,而且对初始聚类中心敏感,对于不同的初始值可能会导致不同的聚类结果。此外k-means算法不适合于发现非凸面形状的类,而且对孤立点和噪声敏感,少量的该类数据能够对平均值产生较大的影响。

针对传统k-means算法的缺点,研究人员提出了一系列改进的k-means算法。文献[2]提出用k-d树结构改进k-means算法;文献[3]用遗传算法优化k-means算法,并将遗传算法的全局搜索能力和k-means算法的局部搜索能力结合在一起;文献[4]提出了一种新的基于数据样本分布选取初始聚类中心的k-means算法。

网络入侵随机性强,事先难以定义类的平均值和类的数目k,偶发事件时有发生。为此,本文提出基于PSO算法的改进k-means(PSO-based k-means)算法。

2 基于PSO的k-means算法

PSO(Particle Swarm Optimization)算法是基于群体智能理论的优化算法,其基本思想是:优化问题的每个解称为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

到稿日期:2010-06-21 返修日期:2010-09-20 本文受江苏省产业技术研究与开发基金,苏发改高技发[2008]106号资助。

傅 涛(1980—),男,博士,主要研究方向为计算机网络;孙亚民(1945—),男,教授,博士生导师,主要研究方向为通信与网络、无线传感器网络等。

函数来衡量粒子距离目标位置的远近，粒子根据自己的飞行经验以及其他粒子的飞行经验动态调整飞行速度和位置，追逐当前的最优粒子，在解空间中搜索最好的目标位置，从而得到所优化问题的最优解[7]。

定义 (1)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 $v_i=(v_{i1},v_{i2},\cdots,v_{in})$；

(2)粒子群中第 i 个粒子在 n 维空间经历过的最好位置 $p_i=(p_{i1},p_{i2},\cdots,p_{in})$；

(3)整个粒子群经历过的最好位置 $g=(g_1,g_2,\cdots,g_n)$；

(4)粒子更新自身速度和位置的公式为：

$v_i=v_i+acc\_const*rand*(p_i-x_i)+acc\_const*rand*(g-x_i)$，$x_i=x_i+v_i$；其中 acc_const 称为加速因子，通常取值为 2，$rand$ 取(0,1)区间的随机数；

(5)适应度函数计算公式为 $f(x_i)=1/d_{max}(x_i,x_j)$，其中 $d_{max}(x_i,x_j)$ 是点 x_i 与聚类中其他所有点 x_j 之间的距离的最大值。

基于 PSO 算法的改进 k-means 算法(PSO-based k-means)描述如下：

输入：包含 n 个对象的数据集、类的数目 k 以及最大迭代次数 MaxIter；

输出：E 收敛或是聚类中心不再变化的 k 个聚类。

(1)将数据集均分为 k 个聚类，每个类称为一个粒子群，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)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，记录数据集中该位置的下标；

(6)根据定义(4)更新每个粒子的速度和位置；

(7)直到达到最大迭代次数 MaxIter，根据 k 个类经历的最好位置的下标输出 k 个初始聚类中心；

(8)从 k 个初始聚类中心出发，运用 k-means 算法进行聚类。

表 1、表 2 分别为 k-means 算法与基于 PSO 的 k-means 的算法聚类精度比较和准则函数 E 比较。采用的数据集是 UCI[8] 数据库的 Iris，Wine 和 Hayes-Roth 3 组数据。

表 1　k-means 算法与基于 PSO 的 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	平均聚类度
k-means	Iris	89.33%	82.00%	87.30%
	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%
	Wine	70.22%	70.22%	70.22%
	Hayes-Roth	80.30%	80.30%	80.30%

由表 1 和表 2 可见，PSO-based k-means 算法对 Iris 和 Hayes-Roth 数据集的聚类精度达到了传统 k-means 的最高聚类精度，分别为 89.33%和 80.30%，其 E 的值等于传统算法的 E 的最小值 78.94084 和 2.17413e4，这说明改进的 k-means 算法对于 Iris 和 Hayes-Roth 数据集有非常好的聚类效果。对于 Wine 数据集，基于 PSO 的 k-means 算法的聚类精度与传统算法的最小精度 70.22%相等，但 E 的值是传统算法的最小值 2.37069e6，这说明基于 PSO 的 k-means 算法对于 Wine 数据集的聚类效果也有一定的改进。

表 2　k-means 算法与基于 PSO 的 k-means 算法准则函数 E 比较

算法	数据集	E_{max}	E_{min}	E_{avg}
k-means	Iris	145.27932	78.94084	98.51058
	Wine	2.64697e6	2.37069e6	2.42393e6
	Hayes-Roth	2.17413e4	2.17179e4	2.17400e4
基于 PSO	Iris	78.94084	78.94084	78.94084
	Wine	2.37069e6	2.37069e6	2.37069e6
	Hayes-Roth	2.17413e4	2.17413e4	2.17413e4

3　PSO-based k-means 算法在入侵检测中的应用

本文将 PSO-based k-means 算法用于入侵检测系统的规则挖掘处理模块，对比分析 k-means 算法和 PSO-based k-means 算法网络入侵检测率和误报率。

测试数据是 KDD Cup1999[10] 数据集中的“kddcup. data_10_percent”，该数据集共有 494020 条记录。其中正常记录条数 97277，其余为入侵记录条数，攻击类型分为 4 类：拒绝服务 DoS(Denial of service)、探测 Probe、远程访问本地 R2L(Remote to Local)和用户访问服务器 U2R(User to Root)。每条数据有 33 个数值型属性和 8 个符号型属性，共提供了 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[9]。

由于每条数据的数值型属性的数值波动范围较大，因此需要对每个属性进行标准化处理。假设所有数据数目为 n，x 为每条数据的任意一项属性，数据预处理的步骤如下：

(1)统计每个符号型属性的数目 m，对每个符号属性赋值，例如，第 i 个属性就赋值 i/m；

(2)按照式(3)计算每个属性的平均值；

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i \tag{3}$$

(3)按照式(4)计算每个属性的标准差：

$$\sigma(x)=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2} \tag{4}$$

(4)按照式(5)标准化每个属性：

$$x_i'=\frac{x_i-\bar{x}}{\sigma(x)} \tag{5}$$

用 k-means 算法进行 20 次实验，取平均值，结果如表 3 所列。用 PSO-based k-means 算法进行实验，最大迭代次数 MaxIter 为 20，实验结果如表 4 所列。

表 3　k-means 算法入侵检测结果

组数	检测率	误报率	准则函数 E
1 组(Dos)	84.9%	12.7%	3.287e9
2 组(Probe)	82.0%	6.5%	3.332e9
3 组(R2L)	78.7%	10.9%	3.394e9
4 组(U2R)	79.7%	8.9%	3.334e9

表 4　PSO-based k-means 算法入侵检测结果

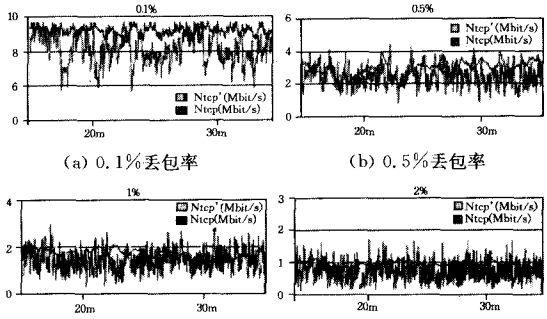
组数	检测率(检测率提高率)	误报率(误报率降低率)	准则函数 E
1 组(Dos)	91.0%(07.2%)	1.9%(84.6%)	3.226e9
2 组(Probe)	97.3%(18.7%)	1.4%(78.6%)	3.232e9
3 组(R2L)	87.0%(10.5%)	3.1%(71.2%)	3.289e9
4 组(U2R)	93.7%(17.6%)	4.1%(53.9%)	3.499e9

从表 3、表 4 可见，PSO-based k-means 算法较传统 k-means 算法在检测率方面有很大的提高，分别提高了 7.2%，

(下转第 73 页)

5.2 Ntcp,Ntcp'带宽估计的性能比较

设瓶颈带宽为10M,往返时延 $Rtt=0.1$s,在丢包率分别为0.1%,0.5%,1%和2%条件下进行Ntcp和Ntcp'带宽估计的仿真,比较结果如图5所示。



(a) 0.1%丢包率 (b) 0.5%丢包率

(c) 1%丢包率 (d) 2%丢包率

图5 Ntcp,Ntcp'带宽估计

由图5可知,不同随机丢包率下,Ntcp'估计带宽的波动比Ntcp算法下小得多,估计的带宽值也较Ntcp高,更接近实际值,说明了Ntcp'改进算法非常有效。

5.3 Reno,Ntcp和Ntcp'的性能

瓶颈带宽为10M,$Rtt=0.1$s时,设置不同的丢包率,对3种算法进行对比仿真,得到的吞吐量如图6所示。

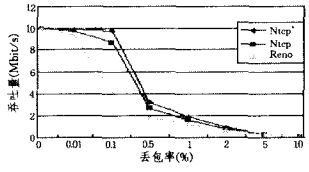


图6 不同丢包率条件下的吞吐量

由图6可以看出同一丢包率时,Ntcp算法下的吞吐量明显高于Reno算法,Ntcp'算法下的吞吐量又较Ntcp高。可以说明Ntcp和Ntcp'算法对具有无线随机丢包的网络环境具有较好的适应性,Ntcp'改进算法在Ntcp的基础上又提高了网络的性能,改进算法是有效的。

结束语 本文在Ntcp算法基础上提出了新的拥塞控制算法。理论分析和仿真结果表明Ntcp算法在无线网络下有较好适应性,而改进后的Ntcp'算法表现出更好的性能。通过对发送端数据量的精确检测,得到网络实际可用带宽,很好地控制了拥塞窗口的大小,提高了带宽的利用率。另外,通过分析推导了Reno,Ntcp和Ntcp' 3种算法下吞吐量与丢包率的关系,发现并实验证明了无论是传统的拥塞控制算法Reno还是本文中Ntcp,Ntcp'算法,在忽略超时丢包情况下,当往返时延一定时,网络的吞吐量仅由丢包率决定。相信,这会对今后TCP拥塞控制算法的性能改进具有重大意义。

参 考 文 献

[1] 周慧斌,周铁军,管贵秋,等. 基于仿真的TCP拥塞控制算法研究[J]. 计算机仿真,2007,24(12):121-124

[2] Allman M, Paxson V, Stevens W R. TCP Congestion Control [C]// RFC 2581. IETF, April 1999

[3] Floyd S, Henderson T. The NewReno Modification to TCP's Fast Recovery Algorithm[C]// RFC 2582. IETF, April 1999

[4] Claudio C, Mario G, Saverio M, et al. TCPWestwood: End-to-End Congestion Control for Wired/Wireless Networks [J]. Wireless Networks, 2002, 8(5): 467-479

[5] Tom G, James M, Phatak D S, et al. A True End-to-End TCP Enhancement Mechanism for Mobile Environments[C]// Proceedings of IEEE INFOCOM. v3. 2000: 1537-1545

[6] Cheng Peng-fu, Liew S C. TCP Veno: TCP Enhancement for Transmission over Wireless Access Networks[J]. IEEE Journal on Selected Areasin Communications, 2003, 21(2): 216-228

[7] 李世银,王秀娟,徐冬,等. 一种基于噪声模型的TCP有效带宽估计方法研究[J]. 系统仿真学报,2008(19):5058-5061

(上接第55页)

18.7%,10.5%和17.6%。而误报率则明显降低,误报率的最大降低率达到84.6%。由此说明,PSO-based k-means算法克服了k-means算法容易陷入局部最优解这一缺点,聚类效果优于k-means算法,从而提升了检测效果。

结束语 k-means算法初始聚类中心随机选择,聚类结果随初始聚类中心的不同而波动,从而导致聚类结果不稳定。本文提出的PSO-based k-means算法使用PSO算法优化生成初始聚类中心,得到的聚类结果全局最优,使得聚类结果不会陷入局部最优解。

在此基础上,将基于PSO的改进k-means算法用于入侵检测系统的规则挖掘处理模块。实验结果表明,PSO-based k-means算法的入侵检测率明显高于传统k-means算法,而误报率则大大低于后者。显然,这对改进入侵检测系统的性能有重要意义。

参 考 文 献

[1] McQueen J. Some methods for classification and analysis of multivariate observations [C] // Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability. 1967

[2] Alsabti K, Ranka S, Singh V. An efficient k-means clustering algorithm[C]// IPPS/SPDP Workshop on High Performance Data Mining. Orlando, Florida, 1998

[3] 陆林华,王波. 一种改进的遗传聚类算法[J]. 计算机工程与应用,2007,43(21):170-172

[4] 曹志宇,张忠林,李元韬. 快速查找初始聚类中心的K_means算法[J]. 兰州交通大学学报,2009,28(6):15-18

[5] Ester M, Kriegel H P, Sander J, et al. A density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise[C] // Proceeding the 2nd International Conference on Knowledge Discovery and Dta Mining(KDD). Portland, 1996

[6] http://archive.ics.uci.edu/ml/datasets.html

[7] 李国辉,李恒峰. 基于内容的音频检索:概念和方法[J]. 小型微型计算机系统,2000,21(1):1173-1177

[8] Vapnik V N. 统计学习理论[M]. 许建华,张学工,译. 北京:电子工业出版社,2004

[9] 程佳. 支持向量机与K-均值聚类融合算法研究[J]. 辽宁师范大学,2008

[10] Richard O, Duda P E, Hart D G, et al. 模式分类(第二版)[M]. 李宏东,姚天翔,等译. 北京:机械工业出版社,2007

[11] 张里,彭小峰. 数据挖掘在网络入侵检测系统中的应用[J]. 重庆工学院学报:自然科学版,2008,22(8):135-138