基于信息熵与 K-means 融合算法的网络入侵检测模型

朱娴睿[1],黄英来[2],王成瑞[3]　(1. 黑龙江省财税信息中心,黑龙江哈尔滨 150001;2. 东北林业大学,黑龙江哈尔滨 150040;3. 黑龙江省农业开发评审中心,黑龙江哈尔滨 150040)

摘要　传统 K-means 算法在初始聚类中心选择时具有较大随机性,是影响聚类分析结果的关键因素。利用信息熵辅助选取聚类中心,提出一种信息熵与 K-means 融合算法,并以此为基础构建一种网络入侵检测模型,除完成异常入侵检测外,可使聚类中心随网络变化而动态更新,提高入侵检测效果。通过对比试验,证明了该方法的可行性及有效性。

关键词　信息熵;K-means 算法;入侵检测
中图分类号　S126;TP393.0　　**文献标识码**　A　　**文章编号**　0517 - 6611(2014)17 - 05671 - 02

Network Intrusion Detection Model Based on Algorithm Combining with Information Entropy and K-means

ZHU Xian-rui et al　(Fiscal and Taxation Information Center of Heilongjiang Province, Harbin, Heilongjiang 150001)

Abstract　Traditional K-means algorithm had randomness in selecting initial cluster center, which was the key factor that influenced the clustering results. Using information entropy to auxiliary select the cluster center, an algorithm combining information entropy with K-means was put forward, and a network intrusion detection model based on the algorithm was constructed, this model can detect the abnormal intrusion and make the cluster center change along with the network changes dynamically, which can improve the intrusion detection effect. Experiment results show that this model is feasible and effective.

Key words　Information entropy; K-means algorithm; Intrusion detection

网络入侵检测是一个从网络中的关键节点收集与网络状况及网络行为相关的数据,对其进行分析以从中发现异常行为特征并提供预警的过程[1-2],以此达到监控网络行为和防御网络入侵的目的。由于入侵行为往往具有较大的不确定性,因此利用聚类分析技术提取数据中隐藏的信息,对识别未知入侵行为具有重要意义。李文华研究了基于模糊 C 均值 FCM 聚类的网络入侵检测模型[3];张国锁等针对 FCM 在处理大数据集时的局限性,提出了改进的 FCM 聚类算法并将其应用于入侵检测[4];罗敏等研究了基于 K-means 聚类算法的无监督入侵检测模型[5];李贺玲针对数据分布不均匀的问题,提出了改进的 K-means 算法并进行了试验分析[6]。上述方法均针对算法在可处理数据大小上进行研究改进,未涉及算法核心部分。为此,笔者针对 K-means 聚类算法,考虑到数据初始簇中心的选取是影响该算法聚类结果的主要因素,研究利用信息熵辅助确定聚类中心,并建立一种基于信息熵与 K-means 融合算法(IE-K-means)的网络入侵检测模型,结果表明,基于此改进算法的入侵检测模型具有良好的入侵检测率。

1　信息熵与 K-means 融合算法

1.1　K-means 算法分析

1.1.1　算法基础。在利用算法进行聚类之前,应首先对数据进行标准化处理,进而给定算法聚类结果的评价标准,用于终止或继续算法的执行过程。

(1)数据对象标准化。

$$S_j = \frac{1}{n}\sum_{i=1}^{n}|x_{ij} - \frac{1}{n}\sum_{i=1}^{n}x_{ij}| \quad (1)$$

$$x'_{ij} = \frac{x_{ij} - \frac{1}{n}\sum_{i=1}^{n}x_{ij}}{S_j} \quad (2)$$

式中,n 为数据对象的个数;x_{ij} 为第 i 个数据对象的第 j 个属性的取值;x'_{ij} 为标准化后的第 i 个数据对象的第 j 个属性的取值。标准化后的数据能够去除量纲对聚类过程的影响。

(2)欧几里德距离。

$$d(i,j) = \sqrt{|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \cdots + |x_{in} - x_{jn}|^2} \quad (3)$$

式中,$d(i,j)$ 为数据对象与其所在簇中心的距离;$i = (x_{i1}, x_{i2}, \cdots, x_{in})$ 和 $j = (x_{j1}, x_{j2}, \cdots, x_{jn})$ 为两个 n 维的数据对象。

(3)评价函数。

$$E = \sum_{i=1}^{k}\sum_{x\in d_i}|x - m_i|^2 \quad (4)$$

$$m_i = \frac{\sum_{x\in d_i} x}{|d_i|} \quad (5)$$

式中,E 为所有数据对象的均方误差之和;d_i 为第 i 个数据簇;x 为数据对象;m_i 为 d_i 中所有数据对象的平均值;$|d_i|$ 为 d_i 中数据对象的个数;k 为划分类的个数。E 值越小,聚类效果越好。

1.1.2　算法流程。利用 K-means 算法进行数据聚类的步骤如下:①指定将生成的聚类个数 k;②选取 k 个数据对象作为初始聚类的中心点;③将原始数据初步划分到 k 个聚类中,并重新计算各数据簇的中心点;④打破上阶段的聚类,并根据欧几里德距离最小为原则,把第 j 个数据对象放入对应聚类中,形成新的聚类,同时计算 E 值;⑤重复步骤④,在新聚类与以前聚类集合相同时,算法结束。

由上可知,算法性能优劣主要受步骤①和②的影响,而聚类个数 k 的确定往往需根据实际情况而定,因此初始聚类中心的确定是影响该算法的关键因素。考虑到聚类中心的选取具有较大随机性[7],该研究利用信息熵来辅助选择聚类中心,进而优化聚类效果。

基金项目　中央高校基本科研业务费专项资金资助(2572014CB25);黑龙江省自然科学基金项目(C201347)。
作者简介　朱娴睿(1978 -),女,黑龙江哈尔滨人,工程师,从事网络安全研究。
收稿日期　2014-05-16

1.2 信息熵 信息熵用于衡量一个随机变量信息的不确定性,其值越大,表明数据越无序;反之,表明数据越有序,数据越相似。若使用信息熵评价聚类效果的优劣,则熵值越小,簇内数据越相似,聚类效果越好。

一个随机变量的信息熵可表示为:

$$E(X) = -\sum_{x \in s(x)} \log_n[p(x)] \quad (6)$$

式中,$S(X)$为 X 的可能取值集合;$p(x)$为 x 的概率函数。

如果 $X = \{x_1, x_2, \cdots, x_n\}$ 包含多个属性,其信息熵可表示为:

$$E(X) = -\sum_{x_1 \in s(x_1)} \sum_{x_2 \in s(x_2)} \cdots \sum_{x_n \in s(x_n)} p(x_1, x_2, \cdots, x_n) \log_n[p(x_1, x_2, \cdots, x_n)] \quad (7)$$

若 $x_1, x_2, \cdots, x_n$ 之间相互独立,则:

$$E(X) = -\sum_{x_1 \in s(x_1)} \sum_{x_2 \in s(x_2)} \cdots \sum_{x_n \in s(x_n)} p(x_1), p(x_2) \cdots p(x_n) \log_n[p(x_1), p(x_2) \cdots p(x_n)] \quad (8)$$

1.3 基于信息熵的改进 K-means 算法 该研究主要研究利用信息熵优化初始聚类中心的选取,降低选择过程的随机性,在改进的基础上进行聚类能够获得更好的聚类结果。

假设样本空间 M 中包含有 n 个记录,计算其中每个记录的信息熵值,然后从第 1 个记录开始,对比当前记录与其他记录的熵值,以最小值为当前记录的基准熵值,其对比矩阵如表 1 所示。

获得基准熵值集 $Base(M) = \{\min E(M_1, M_j), \min E(M_2, M_j), \cdots, \min E(M_n, M_j)\}$,$1 \leqslant j \leqslant n$,对基准熵值由大到小进行排序,得到排序后的基准熵值集 $SortBase(M)$,熵值越大,说明其对应的记录与其他记录越不相似,则比较适合做初始聚类的中心。结合 K-means 算法步骤①确定的聚类个数 k,选取 $SortBase(M)$中前 k 个熵值对应的记录,即为最不相似的前 k 个记录,将其作为初始聚类中心即可。

表 1 信息熵值对比矩阵

i	j					
	1	2	3	…	n	$Base$
1	$E(M_1, M_1)$	$E(M_1, M_2)$	$E(M_1, M_3)$	…	$E(M_1, M_n)$	$\min E(M_1, M_j)$
2	$E(M_2, M_1)$	$E(M_2, M_2)$	$E(M_2, M_3)$	…	$E(M_2, M_n)$	$\min E(M_2, M_j)$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
n	$E(M_n, M_1)$	$E(M_n, M_2)$	$E(M_n, M_3)$	…	$E(M_n, M_n)$	$\min E(M_n, M_j)$

2 基于改进算法的网络入侵检测模型

基于信息熵与 K-means 融合聚类算法,研究一种基于 IE-K-means 算法的网络入侵检测模型,其结构如图 1 所示。由图 1 可知,该模型包含 4 个核心部件,各部分的主要作用如下。

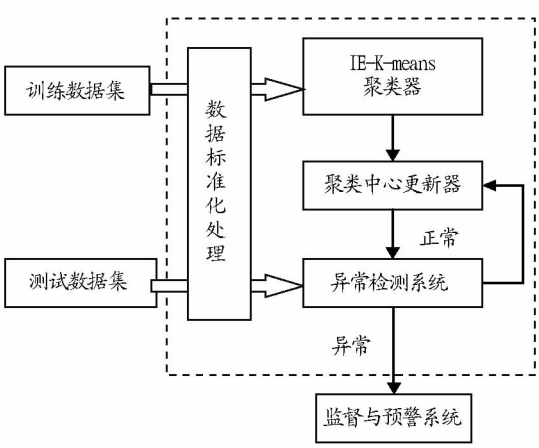

图 1 基于 IE-K-means 的网络入侵检测模型

2.1 数据标准化处理 由于数据可能同时包含数值属性和符号属性,因此,需首先对符号属性进行数值化处理,再与数值属性一起进行标准化处理,统一数据属性的量纲,避免出现大数占优势,小数不能充分发挥作用的情况。标准化处理方法已在"**1.1.1**"中进行阐述。

2.2 IE-K-means 聚类器 该部件是网络入侵检测模型最重要的部分,以信息熵与 K-means 融合算法为核心进行构建,用于训练模型。结合上述的改进算法及 K-means 算法的流程,生成正常类中心集 $NormalC = \{nc_1, nc_2, nc_3, \cdots, nc_n\}$ 和异常类中心集 $AbNormalC = \{anc_1, anc_2, anc_3, \cdots, anc_n\}$,将其保存至聚类中心更新器,用于异常检测。

2.3 异常检测系统 该部件在新的连接到来时,利用式(4)计算连接数据分别与 $NormalC$ 和 $AbNormalC$ 中元素之间的距离,按照最小距离对应的类中心性质标记新连接的属性,一方面将其送入监督与预警系统,实现网络入侵的日常动态监督与报警;另一方面将标记的连接送入聚类中心更新器,用于根据网络变化动态更新类中心集。

2.4 聚类中心更新器 该部件同时接收训练得到的类中心集和异常检测时得到的连接标记结果,实现类中心集随网络变化而动态更新,提高异常检测准确率。若计算得到连接 $X = \{x_1, x_2, \cdots, x_n\}$ 与类中心集 $C = \{c_1, c_2, \cdots, c_n\}$ 距离最近,则将 X 标记为类 C,同时更新类 C,将其类中心修改为连接 X 与类中心对应属性值的均值。

3 仿真试验与分析

为了验证 IE-K-means 算法及基于该算法的网络入侵模型的可行性与有效性,利用 KDDCUP99 数据包进行仿真试验与分析。选择 7 200 条 DoS 攻击数据,其中 5 500 条作为训练数据集,用于训练模型,其余 1 700 条作为测试数据集,用于检测模型入侵检测的有效性。试验分别采用 K-means 算法及 IE-K-means 算法进行对比测试,算法优劣的评价标准为:

$$DetectRate = \frac{detectednum}{totalnum} \times 100\% \quad (9)$$

$$FalseDetectRate = \frac{falsedetectednum}{totalnormalnum} \times 100\% \quad (10)$$

式中,$DetectRate$ 为检测率;$detectednum$ 为检测到的入侵个

(下转第 5682 页)

3 图像抖动分析

在比赛中,对于短跑运动而言,时间是赢得比赛的一项苛刻要求。因此,在运动步态方面,加大了关节电机转动速度,使仿人机器人在保证行进稳定性的前提下运动速度尽量提高到最快。在快速的跑步运动下,仿人机器人的摄像头会随着机身浮动,造成抖动,使得摄像头获取图像变得困难。其中最大的问题就是图像模糊。对于图像模糊的解决方案可以采用去模糊化等一些算法进行实现。但是越复杂的算法往往会需要越多的时间,这对短跑比赛来说是不利的。通过试验,发现机器人调用的系统跑步命令,以最坏偏移角度来进行跑步,所需要的运动时间可以通过试验获得。因此,针对画面抖动这个问题,采用了特殊的处理方式,可以以运动时间为单位,对视觉识别系统的图片获取部分进行屏蔽。即在特定的时间内,视觉系统采集的图像不作为判断偏移的依据,而是先进行跑步运动一段时间之后,再调用停止命令,使摄像头稳定,然后再进行图像的获取与处理。这样,既保证了不消耗太多时间处理图像模糊,又保证了图像采集的清晰化,为仿人机器人的偏移判断提供稳定的依据。

4 仿真与试验

图 2 所示的是仿人机器人采用该研究提出的算法进行短跑运动的实况截图。每张截图的时间间隔为 1 s。试验显示,仿人机器人能够迅速顺利地到达终点目标。

5 结论

该研究设计了一个嵌入视觉识别系统的仿人机器人控制系统。通过采集摄像头图像,获取目标颜色信息,调用系统底层动作库,实现跑步运动。针对 FIRA 比赛的特点,对仿人机器人跑步前进动作进行设计,并通过识别结果进行方向调整。通过试验显示,具有视觉识别系统的仿人机器人具有很好的试验效果。

参考文献

[1] BRUCE J,BALCH T,MANUEL A M. Fast and inexpensive color image segmentation for interactive robots[C]//International conference on robots and system. Takamatsu:IEEE Press,2000:2061 - 2066.

[2] 谭宝成,牟云霞,程智远,等. 全自主移动机器人视觉系统图像分割方法研究[J]. 西安工业大学学报,2007,27(5):471 - 473.

[3] 夏泽洋,陈恳,熊璟,等. 仿人机器人运动规划研究进展[J]. 高技术通讯,2007,17(10):1092 - 1099.

[4] PHILIPP M,CHESTNUTT J,CHUFFER J,et al. Vision - guided humanoid footstep planning for dynamic environments[C]//Proceedings of IEEE/ RAS International Conference on Humanoid Robots. USA,2005:13 - 18.

[5] 魏航信. 仿人跑步机器人快速跑步研究[D]. 西安:西安电子科技大学,2006.

[6] KUFFNER J J,NISHIWAKI K,KAGAMI K,et al. Footstep planning among obstacles for biped robots[C]//Proceedings of IEEE/ RSJ International Conference on Intelligent Robots and Systems. Wailea,Hawaii,2001:500 - 505.

[7] KUFFNER J J,NISHIWAKI K,KAGAMI S,et al. Motion planning for humanoid robots[J]. Transactions in Advanced Robotics,2005,15:365 - 374.

[8] KAVRAKI L,LATOMBE J. Randomized preprocessing of configuration space for fast path planning[C]//IEEE int conf on Robotics and Automation. San Diego:IEEE Press,1994:2138 - 2139.

(上接第 5672 页)

数;*totalnum* 为所有入侵的个数;*FalsedetectRate* 为误警率;*falsedetectednum* 为把正常数据误检为异常数据的个数;*totalnormalnum* 为所有正常数据的个数。

试验采取不同的聚类个数 k 值,首先将训练数据进行聚类,得到聚类中心集,再将测试数据集送入异常检测系统进行入侵检测,并计算各数据集的 *DetectRate* 和 *FalseDetectRate*,对比试验结果如表 2 所示。

表 2 对比试验结果 %

k	K-means 算法		IE-K-means 算法	
	DetectRate	*FalseDetectRate*	*DetectRate*	*FalseDetectRate*
20	85.21	3.10	87.33	0.05
30	87.53	6.64	90.46	0.24
40	95.62	9.86	98.23	0.33

从表 2 可以看出,该研究建立的基于信息熵与 K-means 融合算法的网络入侵检测模型是可行的,且改进算法在不同聚类数目上的检测率及误警率上均优于传统 K-means 算法。

4 结论

该研究基于网络入侵数据的特征和现有入侵检测研究存在的问题,提出了一种基于信息熵与 K-means 融合算法的网络入侵检测模型,结果表明,该模型可行,且较传统 K-means 算法而言,提高了入侵检测的检测率,降低了误警率。但该算法及模型的实现,尚未考虑算法的执行效率问题,下一步应研究在尽可能短的时间内完成入侵检测的实现方法。

参考文献

[1] 陈小辉. 基于数据挖掘算法的入侵检测方法[J]. 计算机工程,2010,36(17):72 - 76.

[2] 李洋. K-means 聚类算法在入侵检测中的应用[J]. 计算机工程,2007,33(14):154 - 156.

[3] 李文华. 基于聚类分析的网络入侵检测模型[J]. 计算机工程,2011,37(17):96 - 98.

[4] 张国锁,周创明,雷英杰. 改进 FCM 聚类算法及其在入侵检测中的应用[J]. 计算机应用,2009,29(5):1336 - 1338.

[5] 罗敏,王丽娜,张焕国. 基于无监督类的入侵检测方法[J]. 电子学报,2003,31(11):1714 - 1716.

[6] 李贺玲. 数据挖掘在网络入侵检测中的应用研究[D]. 长春:吉林大学,2013:26 - 30.

[7] 杜强,孙敏. 基于改进聚类分析算法的入侵检测系统研究[J]. 计算机工程与应用,2011,47(11):106 - 108.