基于分箱统计的FCM算法及其在网络入侵检测中的应用*)

傅 涛[1] 孙文静[2] 孙亚民[1]

(南京理工大学计算机学院 南京 210094)[1] (南京审计学院 南京 210029)[2]

摘 要 使用KDDCup99网络入侵检测数据,对传统的FCM(Fuzzy C-Means)算法进行实验,发现该聚类算法在进行聚类划分和孤立点判断时,存在划分粗略性现象。针对该问题,本文提出使用分箱统计的FCM方法来划分和描述数据集的分布。与原有算法相比,不需要频繁更新聚类中心,同时耗时问题也得到较好的改善。文章最后将特征匹配与基于分箱的FCM算法相结合,协同分析网络连接数据记录。实验结果证明,这种协同检测方法的检测率有明显提高,实时性好,能较好地发现新的攻击类型,便于检测知识库的更新。

关键词 FCM算法,分箱统计,特征匹配,协同检测

FCM Algorithm Based on Box-FCM Statistics and its Application in Network Intrusion Detection

FU Tao[1] SUN Wen-Jing[2] SUN Ya-Min[1]

(College of Computer, Nanjing University of Science & Technology, Nanjing 210094)[1] (Nanjing Audit University, Nanjing 210029)[2]

Abstract This paper carried experiments on the traditional FCM (Fuzzy C Means) algorithm by using KDDCup99 intrusion-detection data of network and found that when clustering division and isolated- point judgments were carried out, phenomenon of rough division was existed in this algorithm. To this question the paper presented that classifying and describing the distribution of data sets by using of the statistics box-FCM methods. Compared with the original algorithm it didn't need to update the clustering center frequently and could resolve time-consuming problems effectively. Finally this paper combined feature matching with box-FCM algorithm so that network connection data records could be coordinated analyzed. Experiments proved that the detection efficiency of this real-time method was improved noticeably and some new intrusion ways could be detected to update the knowledge base.

Keywords FCM algorithm, Based on statistical binning, Statistics, Feature matching, Coordinated analysis

针对不同的应用,人们提出了很多模糊聚类算法,比较典型的有基于相似性关系和模糊关系的方法、基于模糊等价关系的传递闭包方法、基于模糊图论的最大支撑树方法,以及基于数据集的凸分解、动态规划和难以辨识关系方法等[1]。

实际中受到普遍欢迎的是基于目标函数的模糊聚类方法。这种方法的基本思想是把聚类归结成一个带约束的非线性规划问题,通过优化求解获得数据集的模糊划分和聚类。在基于目标函数的聚类算法中模糊C均值(FCM, Fuzzy C-Means)算法的理论最为完善、应用最为广泛。

目前许多研究者将聚类分析应用到入侵检测技术中。本文使用KDDCup99网络入侵检测数据,对FCM算法进行实验,发现该聚类算法在进行聚类划分和孤立点判断时,存在划分粗略性现象。针对此问题,本文提出使用分箱统计的FCM方法来划分和描述数据集的分布,与原有算法相比,不需要频繁更新聚类中心,同时耗时问题也得到较好的改善。文章最后将特征匹配与基于分箱的FCM算法相结合,协同分析网络连接数据记录。实验结果证明,这种协同检测方法的检测率有明显提高,实时性好,能较好地发现新的攻击类型,便于检测知识库的更新[2~4]。

1 FCM算法

给定聚类类别数 c, $2\leqslant c\leqslant n$, n 是数据个数,设定迭代停止阈值 ε,初始化聚类原型模式 $P(0)$,设置迭代计数器 $b=0$。

步骤1:用式(1)计算或更新划分矩阵 $U^{(b)}$,对于 $\forall i,k$,如果 $\exists d_{ik}^{(b)}>0$,则有

$$\mu_{ik}^{(b)}=\left\{\sum_{j=1}^{c}\left[\left(\frac{d_{ik}^{(b)}}{d_{jk}^{(b)}}\right)^{\frac{2}{m-1}}\right]\right\}^{-1} \tag{1}$$

如果 $\exists i,r$,使得 $d_{ir}^{(b)}=0$,则有

$\mu_{ir}^{(b)}=1$,且对 $j\neq r,\mu_{ij}^{(b)}=0$

步骤2:用式(2)更新聚类原型模式矩阵 $P^{(b+1)}$:

$$P_i^{(b+1)}=\frac{\sum_{k=1}^{n}(\mu_{ik}^{(b+1)})^m\cdot x_k}{\sum_{k=1}^{n}(\mu_{ik}^{(b+1)})^m},i=1,2,\cdots,c \tag{2}$$

步骤3:如果 $\|P^{(b)}-P^{(b+1)}\|<\varepsilon$,则算法停止并输出划分矩阵 U 和聚类原型 P,否则令 $b=b+1$,转向步骤1。其中 $\|\cdot\|$ 为某种合适的矩阵范数。

由以上算法可看出,整个计算过程就是反复修改聚类中心和分类矩阵的过程,因此常称这种方法为动态聚类或者逐步聚类法。

FCM算法是目前最受欢迎的模糊聚类算法之一,在很多领域已经获得了成功的应用,但该算法在用于入侵检测时还存在一些问题亟待解决。如

(1)模糊加权指数 m 是Bezdek对模糊聚类准则函数WGSS推广时引入的参数,实践中发现,合适的 m 值具有抑

*)基金项目:江苏省产业技术研究与开发基金,苏发改高技发[2006]1106号。傅 涛 博士生,研究方向为入侵检测;孙文静 博士生,研究方向为移动IPV6的安全。

制噪声、平滑隶属函数等功能，但如何优选参数 m，尚缺乏理论指导。尽管存在一些经验值或经验范围，但没有面向问题的优选方法，也缺少参数 m 有效性的评价准则。应用于入侵检测时多数将 m 设置为 2。

(2)尽管模糊聚类是一种无监督的分类，但目前的 FCM 算法却要求聚类原型参数的先验知识(原型的数目及类型)，否则算法会产生误导，从而破坏了算法的无监督性和自动化。

(3)由于模糊聚类函数是非凸的，而 FCM 聚类算法又是迭代爬山，因此很容易陷入局部极值点或鞍点而得不到最优解甚至满意解，同时，大数据量下耗时问题也是困扰人们的一大难题。

2 分箱统计

顾名思义，按某种标准尺度将数据分放于不同的箱中，并为每个分箱打上标记，以标识该分箱中所放数据的特点。在数据量不大时，可以将每个数据都存放在分箱中储存起来。但面对海量网络连接数据，如果仍然将数据存储于分箱中，将占用很大的存储空间。因此这里分箱存储的仅仅是某数据子集的特点。

为存储分箱建立数据库 boxes_of_clusters。在该数据库中有两类表：

(1) 记录聚类的聚内距离分布情况的分箱表。分箱表由 5 个数据项组成，如表 1 所示。每个聚类的聚内距离的分布情况可用多个分箱来描述。

表 1 聚内距离分布情况分箱表

分箱的编号	分箱的底部	分箱的顶部	分箱所占比例权重	分箱的均值
1-n	min	max	proportion	avg

(2)各聚类进行分箱时的比例刻度，其结构如表 2 所示。对于不同的聚类，数据点间的稀疏密度不同，从而聚内距离值的分布密度也不一样。因此，对于不同的聚类在分箱时采用不同的比例刻度来分箱处理。

表 2 各聚类分箱比例刻度表

编号	聚类名称	分箱的比例刻度	更新参数
num	type	scale	update

3 基于分箱统计的 FCM 算法

传统的 FCM 算法是一种无指导无监督性的划分聚类的方法，且很容易陷入局部极值点或鞍点而得不到最优解甚至满意解。同时，在处理网络连接数据记录这样的大数据量时，需要频繁更新聚类中心，算法耗时。本文提出基于分箱统计的 FCM 算法，根据带标识的已知聚类的划分，进行新数据记录的类型判断，不仅可以避免陷入局部极值点或鞍点，而且还可以根据分箱判断是否更新聚类中心，从而避免原来需要频繁更新聚类中心的问题，相对提高处理数据的速度。

传统 FCM 算法中使用隶属函数 μ_{ki} 表示数据记录 x_k 与聚类子集 $X_i(1\leqslant i\leqslant c)$ 的隶属关系，使用最大隶属原则来判断数据记录的归属。这里使用如下的模糊近似度方法进行判断。

$$T_i=\left|1-\frac{d_i}{D_i}\right| \tag{3}$$

对已知的各聚类计算 T_i，若 $\min\{T_i\}$ 在对应聚类的门限范围内，将 $\min\{T_i\}$ 所对应的聚类作为新数据记录所属类；若 $\min\{T_i\}$ 不在对应聚类的门限范围内，则作为新类处理。

处理中当检测出某新的网络连接数据记录属于哪种聚类时，不再像以往的聚类方法立即进行聚类中心的更新，而是将该数据记录到所属聚类的聚类中心的距离 d_i，与该聚类的分箱进行比较，当分箱需要更新时，才对该聚类的聚类中心进行更新。这种方法避免了以往聚类方法中需要频繁更新聚类中心的问题。为便于分箱更新，减少需要更新的数据，对上一节中的分箱结构做了改进，如表 3 和表 4 所示。

表 3 聚内距离分布情况分箱表

分箱的编号	分箱的底部	分箱的顶部	分箱容量	分箱的均值
1-n	min	max	sum	avg

表 4 各聚类分箱比例刻度表

编号	聚类名称	分箱的比例刻度	总容量
num	type	scale	totality

对于不同的聚类，其分箱划分时有的是连续的，有的存在断层，且断层的幅度有大有小。对于连续的分箱段，仅在分箱的最末处存在需要更新的情况，如图 1 所示；对于有断层的分箱段，则在出现断层的分箱的底部或顶部存在需要更新的情况。在分箱的顶部或底部更新的同时，需要更新分箱的均值和容量，以及该聚类的更新参数，如图 2 所示。

图 1 连续型分箱

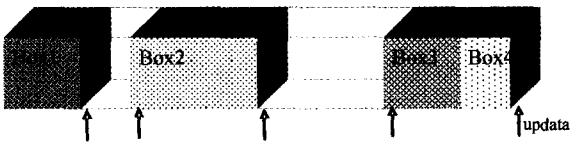

图 2 非连续型分箱

当某数据记录到聚类中心的距离 d_{new} 落在该聚类的某分箱中时，分两种情况讨论：

①当 $d_{new}<\min-scale$ 时，该分箱的底部更新为 $\min_{new}=d_{new}$，使用式(4)更新该分箱的均值。

②当 $d_{new}>\max+scale$ 时，该分箱的顶部更新为 $\max_{new}=d_{new}$，使用式(4)更新该分箱的均值。

$$avg_{new}=\frac{avg_{old}S_{old}+d_{new}}{T_{old}+1} \tag{4}$$

其中参数 S 为聚内距离分布情况分箱表中的分箱容量，参数 T 为分箱比例刻度表中的该聚类的分箱总容量 totality。

$$\text{分箱容量}\quad S_{new}=S_{old}+1 \tag{5}$$

$$\text{分箱总容量}\quad T_{new}=T_{old}+1 \tag{6}$$

聚类中心由离散型属性向量和连续型属性向量两部分组成。对于离散型属性向量的更新分两种情况：

① 新数据记录的某离散型属性向量 m 的离散值 i 在已有的离散值集 I 中，即 $i\in I$。这时，对于该离散型属性向量 m 的离散值 i 的原概率统计值 p_i 的更新如式(7)所示。

$$p'_i=\frac{p_i T_{old}+1}{T_{old}+1} \tag{7}$$

该离散型属性向量 m 的其他离散值 j 的原概率统计值 p_j 的更新如式(8)所示。

$$p'_j=\frac{p_j T_{old}}{T_{old}+1} \tag{8}$$

②新数据记录的某离散型属性向量 m 的离散值 i 不在已有的离散值集 I 中，即 $i\notin I$。对于离散型属性向量 m 的新离散值 i 的概率统计值 p_i 如式(9)所示。

$$p_i=\frac{1}{1+T_{old}} \tag{9}$$

否则，该离散型属性向量 m 已有的离散值 j 的原概率统计值 p_j 的更新如式(10)所示。

$$p'_j=\frac{p_j}{1+T_{old}} \tag{10}$$

对于连续型属性向量中心的更新如式(11)所示。

$$Y'_p=\frac{Y_p T_{old}+X_p}{1+T_{old}} \tag{11}$$

其中 Y_p 为某聚类中心 C 的连续型属性向量值，$11\leqslant p\leqslant 32$，$X_p$ 为某数据的连续型属性向量值。

如果新的网络连接数据记录不属于已有的任何聚类，则根据该数据记录设置一新的聚类，同时建立相应的分箱。

4 基于分箱统计的 FCM 算法实验分析

4.1 实验数据

使用 KDDCup99[5] 的网络入侵检测数据集，是从一个模拟的美国空军局域网上采集来的 9 个星期的网络连接数据，分成具有标识的训练数据和未加标识的测试数据。测试数据和训练数据有着不同的概率分布，测试数据包含了一些未出现在训练数据中的攻击类型，这使得入侵检测更具有现实性。

训练数据集中包含了 1 种正常的标识类型 normal 和 22 种训练攻击类型，如表 5 所示。另外有 14 种攻击仅出现在测试数据集中。

KDDCup99 训练数据集中每个连接记录包含了 41 个固定的特征属性和 1 个类标识，标识用来表示该条连接记录或者是正常的，或者是某个具体的攻击类型。在 41 个固定的特征属性中，9 个特征属性为离散(symbolic)型，其他均为连续(continuous)型。

通过对 41 个固定特征属性的分析，比较能体现出状态变化的是前 31 个特征属性，其中 9 个离散型，22 个连续型。因此对连接记录的分析处理是针对该 31 个特征属性的。

表 5 KDDCup99 入侵检测实验数据的标识类型

标识类型	含义	具体分类标识
Normal	正常记录	Normal
DOS	拒绝服务攻击	back、land、neptune、pod、smurf、teardrop
Probing	监视和其他探测活动	ipsweep、nmap、portsweep、satan
R2L	来自远程机器的非法访问	ftp_write、guess_passwd、imap、multihop、phf、spy、warezclient、warezmaster
U2R	普通用户对本地超级用户特权的非法访问	buffer_overflow、loadmodule、perl、rootkit

本文实验使用 KDDCup99 中的网络入侵检测数据包 kddcup_data_10percent。该数据包是对 kddcup_data 数据包(约 490 万条数据记录)10%的抽样。10%抽样的结果是仅仅减少了 normal、neptune、smurf、ipsweep、nmap、portsweep 和 satan 这七个类型记录的个数，对所包含的记录类型总个数并没有改变。

4.2 基于分箱统计的 FCM 算法实验分析

在代理端的检测中使用分箱统计的 FCM 算法，对 KDDCup99 网络入侵检测数据包中的测试数据包进行实验，检测结果如表 6 所示，对于 back，land，pod，nmap 攻击类型检测结果理想，其余类型检测结果均不理想。这说明使用基于分箱统计的 FCM 算法检测时，只能部分解决离散型属性向量关联规则在几个聚类中都存在时的问题。

表 6 基于分箱统计的 FCM 算法下测试数据包的测试结果

类型	具体类型	测试包中数据量	正确检测数据量	检测率
DOS	back	1098	1095	99.7%
	land	9	9	100%
	neptune	58001	19127	33.0%
	pod	87	87	100%
	smurf	164091	131161	79.9%
	teardrop	12	8	66.7%
Probing	ipsweep	306	0	0%
	nmap	84	80	95.2%
	portsweep	354	323	91.2%
	satan	1633	1149	70.4%
R2L	ftp_write	3	1	33.3%
	guess_passwd	4367	32	0.7%
	imap	1	0	0%
	multihop	18	3	16.7%
	phf	2	1	50%
	spy	无		
	warezclient			
	warezmaster	1602	0	0%
U2R	buffer_overflow	22	10	45.5%
	Loadmodule	2	0	0%
	Perl	2	1	50%
	rootkit	13	4	30.8%
	normal	60593	2871	4.7%

检测中出现判断错误的原因可能有以下三种：

(1)离散型属性向量关联规则在多个聚类中存在时，往往将相关的数据记录判为关联规则对应的离散型属性向量概率高的聚类。

以测试数据包中 ipsweep 攻击记录为例，ipsweep 攻击记录有 306 条，检测中 300 条记录被判断为 nmap 类型。这 300 条记录的离散型属性向量均相同，但所组成的关联规则在 nmap 类型中也有，且 nmap 中各离散型属性向量的概率值高于 ipsweep 类型。因此根据基于分箱统计的 FCM 算法判断，这 300 条记录为 nmap 型。

(2)离散型属性向量关联规则存在于多个聚类中，其中某聚类的概率值较高，这时与该关联规则相关的聚类的部分数据记录可以正确检测出来，其余部分检测判断错误。

以 neptune 类型为例，检测数据包中有 58001 条记录，仅 19127 条记录检测判断正确。对判断错误的数据记录进行了分析，结果表明绝大部分离散型向量的关联规则与检测判断的数据记录相同。其原因在于 neptune 类型中的大部分数据记录到 neptune 聚类中心的距离分布在 neptune 整个分箱的下部，因此使用模糊近似度 T_i 进行判断时，会将这部分数据记录判断为属于其他攻击类型。

(3)测试数据包中某数据记录的离散型属性向量关联规则,不在训练数据包所得的离散型属性向量关联规则中。

对此,本文实验研究中将出现概率较高的离散型属性向量关联规则,附加上部分连续型属性规则。例如,某离散型属性向量关联规则只在正常网络连接数据记录中出现,具有独有性,而实际情况中不断出现的新攻击类型,总是会违反正常网络连接数据记录的某些特性。基于此,若某网络连接数据记录的离散型属性向量关联规则与正常离散型属性向量关联规则相同,但违反某正常连续型属性向量规则,且不在已知攻击类型的相应连续型属性向量规则范围内,则该数据记录为新的攻击类型,或属于某已知攻击类型。

4.3 特征匹配与聚类算法的协同检测

表7 特征匹配与聚类算法协同检测时测试数据包的检测结果

类型	具体类型	测试包中数据量	正确检测数据量	检测率
DOS	back	1098	1098	100%
	land	9	9	100%
	neptune	58001	52280	90.1%
	pod	87	87	100%
	smurf	164091	163772	99.8%
	teardrop	12	10	83.3%
Probing	ipsweep	306	253	82.7%
	nmap	84	80	95.2%
	portsweep	354	286	80.8%
	satan	1633	1312	80.3%
R2L	ftp_write	3	2	66.7%
	guess_passwd	4367	3673	84.1%
	imap	1	1	100%
	multihop	18	13	72.2%
	phf	2	1	50%
	spy	无		
	warezclient			
	warezmaster	1602	1063	66.4%
U2R	buffer_overflow	22	12	54.5%
	loadmodule	2	2	100%
	perl	2	1	50%
	rootkit	13	10	76.9%
normal		60593	60223	99.4%

将连续型属性规则的特征匹配,与基于分箱统计的FCM算法结合,其检测过程分以下两种情况:

(1)当新数据记录的离散型属性向量组成的离散型关联规则在离散型关联规则集中为独有时,计算该新数据记录到此离散型关联规则对应聚类的距离,使用式(3)判断新数据记录是否属于该聚类。

①如果属于该聚类,则与该聚类的分箱比较,判断是否要更新该聚类。

②如果不属于该聚类,则根据该新数据记录建立一新的聚类。

(2)当新数据记录的离散型属性向量组成的离散型关联规则在离散型关联规则集中为非独有时,与有该规则的聚类的连续型规则比较,判断在哪个聚类的连续型规则范围内。

①当只在一个聚类的连续型规则内时,该新数据记录属于该聚类,并计算到该聚类的距离,判断是否要更新该聚类。

②当在几个聚类的连续型关规则内时,计算到这些聚类的距离,使用式(3)判断新数据记录属于哪个聚类,并与该聚类的分箱比较,判断是否要更新该聚类。

③当不在任何已知聚类的连续型关规则内时,该新数据记录为新的类型,为此数据记录建立一新的聚类。

对测试数据包进行实验,其结果如表7所示,总检测率达97.2%。实验证明检测引擎上使用特征匹配与聚类算法的协同分析检测,不但速度快,正确性较高,且便于自动更新检测模块,发现新的攻击类型。

结束语 在网络入侵检测数据实验中,基于分箱统计的FCM聚类算法对部分聚类的检测率较为理想,但整体检测效果并不好。其原因可能在于:离散型属性向量关联规则在多个聚类中存在时,往往将相关的数据记录判为关联规则对应的离散型属性向量概率高的聚类;当某聚类的概率值较高时,与该关联规则相关的聚类的部分数据记录可以正确检测出来,其余部分检测判断错误;测试数据包中某数据记录的离散型属性向量关联规则,不在训练数据包所得的离散型属性向量关联规则中。

将特征匹配与基于分箱的FCM算法相结合,协同分析网络连接数据记录,这时检测率有明显提高,实时性好,能较好地发现新的攻击类型,便于检测知识库的更新。

参考文献

1 高新波著.模糊聚类分析及其应用.西安:西安电子科技大学出版社,2004.1

2 Ramaswamy S,Rastogi R,Shim K. Efficient algorithms for mining outliers from large data sets. In: Proceedings of the ACM SIGMOD International Conference on Management of Data, Dallas, TX, USA, 2000. 427~438

3 Portnoy L,Eskin E,Stolfo S J. Intrusion detection withunlabeled data using clustering. In: Proceedings of theACM Workshop on Data Mining Applied to Security,Philadelphia, PA, 2001

4 Sequeira K, Zaki M. ADMIT: Anomaly-based data mining for intrusions. In: Proceedings of the 8th ACMSIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, Alberta, Canada, 2002. 386~395

5 http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.htm

(上接第15页)

的分类正确率,这表明了该算法可以在较大规模的数据集上进行有效的特征选择。

参考文献

1 Pawlak Z. Rough Sets: Theoretical Aspects of Reasoning about Data. Kluwer Academic publishers,1991

2 Liu Q. Granules and Application of Granular Computing in Logical Reasoning. Chinese of Computer Research and Development, 2004, 41(4):546~551

3 Yao Y Y. A Partition Model of Granular Computing. T Rough Sets,2004.232~253

4 李道国,苗夺谦,张红云.粒度计算的理论、模型与方法.复旦学报, 2004, 43(5): 837~841

5 Yao Y Y,Yao J T. Granular computing as a basis for consistent classification problems. In: Proceedings of PAKDD'02 Workshop on Toward the Foundation of Data Mining, Taipei,2002. 101~106

6 邓蔚,王国胤,吴渝.粒计算综述.计算机科学,2004, 31(z2):178~181

7 Yao Y Y. Three Perspectives of Granular Computing. 南昌工程学院学报,2006,25(2): 16~20

8 Ding C, Peng H C. Minimum Redundancy Feature Selection from Microarray Gene Expression Data. In: Proceedings of the Computational Systems Bioinformatics , California ,2003