一种进化半监督式模糊聚类的入侵检测算法

杨晓强[1,2]
YANG Xiao-qiang[1,2]

1.西安电子科技大学 微电子学院,西安 710071
2.西安科技大学 计算机系,西安 710054
1.School of Microelectronics,Xidian University,Xi'an 710071,China
2.Department of Computer,Xi'an University of Science and Technology,Xi'an 710054,China
E-mail:xyangq@sina.com

YANG Xiao-qiang.Algorithm for intrusion detection based on evolutionary semi-supervised fuzzy clustering.Computer Engineering and Applications,2008,44(4):33-35.

Abstract: An algorithm for intrusion detection based on evolutionary semi-supervised fuzzy clustering is proposed which is suited for situation that gaining labeled data is more difficulty than unlabeled data in intrusion detection systems.The algorithm requires a small number of labeled data only and a large number of unlabeled dada,and class labels information provided by labeled data is used to guide the evolution process of each fuzzy partition on unlabeled data,which plays the role of chromosome.This algorithm can deal with fuzzy label,uneasily plunges locally optima and is suited to implement on parallel architecture. Experiments show that the algorithm can improve classification accuracy and has high detection efficiency.
Key words: intrsion detection;sem-supervised learning;clustering;evolutionary programming

摘 要:在入侵检测系统中,未知标签数据容易获得,标签数据较难获得,对此提出了一种基于进化半监督式模糊聚类入侵检测算法。算法利用标签数据信息担任染色体的角色,引导非标签数据每个模糊分类的进化过程,能够使用少量的标签数据和大量未知标签数据生成入侵检测系统分类器,可处理模糊类标签,不易陷入局部最优,适合并行结构的实现。实验结果表明,算法有较高的检测率。
关键词:入侵检测;半监督学习;聚类;进化规划
文章编号:1002-8331(2008)04-0033-03 **文献标识码:**A **中图分类号:**TP393

1 引言

入侵检测作为一种积极主动的网络安全防护技术,目的是为系统提供实时发现入侵行为并及时采取相应防护手段。它具体包括数据收集、行为分类、报告错误和响应反击等方面,其中用到的数据可以由专门的网络管理系统(NMS)或从网络和系统的日志文件中得到, 而数据推导和数据分类是其中的核心。数据分类是定义攻击和识别攻击的过程,具体实现这个过程的技术较多,如模式匹配、统计分析、完整性分析等方法,其本质大多是比较正常状态和考察状态之间的差异,以此来判断系统是否受到了入侵[1]。

入侵检测系统和大多数机器学习系统一样,要依赖于已有的标签数据。标签数据获取困难,它需要专业人员花费大量时间去收集和识别。对于复杂的学习任务,提供足够量的标签数据变得不太现实。获得未知标签数据比标签数据要容易得多,但只用这些未知标签数据去构造分类器效果较差[2]。例如:聚类技术可用于识别未知标签数据的自然群,但类的划分不总是与数据的自然聚类一致。本文将进化半监督式模糊聚类算法(Evolutionary Semi-Supervised Fuzzy Clustering,ESSFC)[3]用于入侵检测系统,标签数据担任染色体的角色,提供的类标签信息用于引导每条染色体的进化过程,每条染色体的适应度用未知标签数据的模糊聚类差异和标签数据的错分类误差联合评估。获得的分类结构用来对未来新数据进行分类。

2 进化半监督模糊聚类入侵检测算法

获取的少量标签数据和大量未知标签数据,每个数据在一些合适的特征空间可以描述成一个特征矢量。所获取的标签数据和未知标签数据可以表示为矩阵的形式:

$$\boldsymbol{X}=\lfloor \overset{l}{x}_1,\cdots,\overset{l}{x}_{n_l},\cdots,\overset{u}{x}_1,\cdots,\overset{u}{x}_{n_u}\rfloor \quad (1)$$

这里下标 l 表示标签数据,下标 u 表示未知标签数据,由式(1)导出一个模糊 $\boldsymbol{C}$ 矩阵表示如式(2):

基金项目:国家自然科学基金(the National Natural Science Foundation of China under Grant No.90607008)。
作者简介:杨晓强(1971-),男,讲师,博士研究生,主要从事计算机网络安全及集成电路设计方面的研究。
收稿日期:2007-07-23 **修回日期:**2007-11-13

$$U=\begin{bmatrix} \overbrace{\underbrace{u_{11}^{l}}_{\text{标签样本1}} \cdots \underbrace{u_{1n_l}^{l}}_{\text{标签样本}n_l}}^{u^l} & \overbrace{\underbrace{u_{11}^{u}}_{\text{未知标签样本1}} \cdots \underbrace{u_{1n_u}^{u}}_{\text{未知标签样本}n_u}}^{u^u} \\ u_{21}^{l} \cdots u_{2n_l}^{l} & u_{21}^{u} \cdots u_{2n_u}^{u} \\ \vdots \quad\quad \vdots & \vdots \quad\quad \vdots \\ u_{c1}^{l} \cdots u_{cn_l}^{l} & u_{c1}^{u} \cdots u_{cn_u}^{u} \end{bmatrix} \tag{2}$$

在该矩阵中，u_{ih}^{l} 和 u_{ij}^{u} 应该满足条件：

$$u_{ih}^{l}\in[0,1] \quad \sum_{i=1}^{c}u_{ih}^{l}=1 \quad 1\leqslant i\leqslant c \quad 1\leqslant h\leqslant n_i \tag{3}$$

$$u_{ij}^{u}\in[0,1] \quad \sum_{i=1}^{c}u_{ij}^{u}=1 \quad 1\leqslant i\leqslant c \quad 1\leqslant j\leqslant n_u \tag{4}$$

该问题目标是用 $\boldsymbol{X}$ 构造一个分类器，用尽可能小的误差将一个未来的新模式分配到一个或多个预先定义的类里。主要思想是在 $\boldsymbol{X}^u$ 上发现一个模糊 $\boldsymbol{C}$ 划分，它可以同时使未知标签数据的聚类差异和标签数据的错分类误差最小化，然后用获得的分类器去分类未来的新数据。

2.1 标签数据的错分类误差

为了获得好的泛化性能，构造的分类器应该是使标签数据的错分类误差最小。算法用标签数据的聚类成员差异测量错分类误差。设在 $\boldsymbol{X}^u$ 上的模糊 $\boldsymbol{C}$ 分类，$\boldsymbol{C}$ 个聚类中心为 $v_1,v_2,\cdots,v_c$ 可以通过式(5)计算：

$$v_i=\frac{\sum_{k=1}^{n_l}(u_{ik}^{l})^2x_k^{l}+\sum_{k=1}^{n_u}(u_{ik}^{u})^2x_k^{u}}{\sum_{k=1}^{n_l}(u_{ik}^{l})^2+\sum_{k=1}^{n_u}(u_{ik}^{u})^2} \tag{5}$$

标签数据的模糊成员关系可通过式(6)重新计算，其中 $i=1,2,\cdots,c$，$j=1,2,\cdots,n_l$。

$$u_{ik}^{l'}=\left[\sum_{i=1}^{c}\left(\frac{\|x_j^{l}-v_i\|_c^2}{\|x_j^{l}-v_h\|_c^2}\right)\right]^{-1} \tag{6}$$

这里，$\boldsymbol{C}$ 是协方差矩阵，$\|x_j^{l}-v_i\|_c^2=(x_j^{l}-v_i)^{\mathrm{T}}C(x_j^{l}-v_i)$。将标签数据的错分类误差表示为 E，E 用于测量矢量 u_{ij}^{l} 和 $u_{ij}^{l'}$ 之间距离的权重和。权重为：$\|x_j^{l}-v_i\|_c^2$，E 的表达式为：

$$E=\sum_{j=1}^{n_l}\sum_{i=1}^{c}(u_{ij}^{l'}-u_{ij}^{l})^2\|x_j^{l}-v_i\|_c^2 \tag{7}$$

2.2 未知标签数据的模糊聚类差异

虽然最小化标签数据的错分类误差对于分类器获得好的泛化能力是十分重要的，但只是最小化标签数据的错分类误差还不够，因为少量标签数据的最小错分类误差很可能导致所谓的过匹配问题，因此需要加入未知标签数据的模糊聚类差异这个分项。模糊聚类差异定义如式(8)：

$$J=\sum_{j=1}^{n_u}\sum_{i=1}^{c}(u_{ij}^{u})^2\|x_j^{u}-v_i\|_c^2 \tag{8}$$

聚类差异的最小化等价于相同类中的数据相似性的最大化。因此说未知标签数据的模糊聚类差异在该问题中可以承担一个能力控制的角色。基于式(7)和(8)，可以定义目标函数式(9)：

$$f(U^u,v)=J+\alpha\cdot E \tag{9}$$

其中，$\alpha>0$ 是调整参数，它用来保持未知标签数据模糊聚类差异和标签数据错分类误差之间的平衡。α 的选择依赖于标签数据集和未知标签数据集的相对尺寸。α 的值应该近似地等于式(9)中两项的权重，一般设置为 n_u/n_l。

问题转化为求式(9)目标函数的最小值问题。有多种方法可以解决这个优化问题，例如：基于微积分的优化算法、遗传算法、半监督模糊聚类算法 Nelder-Mead、单纯形算法等[6]。然而，这些方法可能会陷入局部最优或对初始值非常敏感。进化规划算法不但可以避免局部最优，而且对初始值不敏感[7]。进化规划算法对并行系统结构没有什么限制和要求，适合在目前所有的并行或分布式系统上进行并行处理，可以提高运算速度，因此，本文选用进化规划优化目标函数。

2.3 分类器建立

初始化部分：

随机选取 c 条标签数据，根据数据的标签值在每一类中任意选取一条数据，设置为初始的 c 个聚类中心点：$v_i(1\leqslant i\leqslant c)$。对于标签数据的 $u_{ij}^{l}(i=1,2,\cdots,c)$，可以根据其数据的属性标签确定，所以生成的隶属度矩阵的元素非 0 即 1。获得的 u_{ij}^{l}，满足 $u_{ij}^{l}\in[0,1]$；$\sum_{i=1}^{l}u_{ij}^{l}=1,1\leqslant i\leqslant c,1\leqslant j\leqslant n_l$。

假设种群尺寸为 p，则第 $k(1\leqslant k\leqslant p)$ 条染色体将通过以下步骤生成。

(1)随机产生 c 个实数 $r_{1j},r_{2j},\cdots,r_{cj}(r_{ij}\in[0,1],1\leqslant i\leqslant c)$，作为一条染色体的第 j 个点，$1\leqslant j\leqslant n_u$。

(2)计算 $u_{ij}^{uk}=\frac{r_{ij}}{r_{1j}+r_{2j}+\cdots+r_{cj}}$，$1\leqslant i\leqslant c$，$u_{ij}^{uk}$ 应该满足式(4)。

(3)重复步骤(1)和(2)n_u 次。即 $j=1,2,\cdots,n_u$，产生一条染色体。

(4)重复步骤(1)~(3)p 次，产生初始的 p 条染色体。通过式(10)计算聚类中心 v_i^{k}

$$v_i^{k}=\frac{\sum_{j=1}^{n_l}(u_{ij}^{lk})^2x_j^{l}}{\sum_{j=1}^{n_l}(u_{ij}^{lk})^2} \quad k=1,2,\cdots,p;i=1,2,\cdots,c \tag{10}$$

并计算各自的目标函数值：

$$f^{k}=j^{k}+\alpha\cdot E^{k} \tag{11}$$

下面为进化半监督式模糊聚类算法的主体重复部分。设置后代计数器 $gen=0$，最大后代数为 max_gen。其步骤如下：

(1)用式(12)产生后代 $U^{U(P+K)}$，其中 $k=1,2,\cdots,p;i=1,2,\cdots,c;j=1,2,\cdots,n_u$。

$$u_{ij}^{u(p+j)}=\frac{(u_{ij}^{uk})^2e^{-\|x_j^{u}-v_i^{k}\|_c}}{\sum_{h=1}^{c}(u_{ij}^{uk})^2e^{-\|x_j^{u}-v_h^{k}\|_c}} \tag{12}$$

(2)用式(13)和(14)计算新的聚类中心 v_i^{p+k} 和新的目标函数 f^{p+k}。

$$v_i^{p+k}=\frac{\sum_{j=1}^{n_l}(u_{ij}^{lk})^2x_j^{l}+\sum_{j=1}^{n_u}(u_{ij}^{uk})^2x_j^{u}}{\sum_{j=1}^{n_l}(u_{ij}^{lk})^2+\sum_{j=1}^{n_u}(u_{ij}^{uk})^2} \tag{13}$$

$$f^{p+k}=J^{p+k}+\alpha\cdot E^{p+k} \tag{14}$$

(3)根据 f 的值,从 $p+p$ 矩阵中选择最合适的 p 个(即:目标函数值 f 小的 p 个),形成下一代的种群。

(4)$gen=gen+1$。

(5)重复步骤(1)~(4),直到 max 为 max_gen,或者超时,循环结束。

循环结束后,得到分类器模型,可以用来对未知数据进行分类。

2.4 分类器进行新的未知数据分类

设给定一个新的数据 x,通过 ESSFC 算法获得的 c 个聚类中心点:$v_i(1\leqslant i\leqslant c)$,则 x 属于第 i 类的可能性为 u_i,可以用式(15)计算。

$$u_i=\left[\sum_{h=1}^{c}\left(\frac{\|x-v_i\|_c^2}{\|x-v_h\|_c^2}\right)\right]^{-1} \tag{15}$$

x 可根据对各类的隶属度值,然后取 c 个 u_i 中最大的一个设为 u_k,则数据 x 属于 k 类。

3 实验仿真

实验使用的入侵数据是模拟的 KDDCUP99[7]入侵数据集中的数据。该数据集共提供了大约 4 900 000 条数据,每条数据有 42 维属性,包括 38 个数字型特征,3 个字符型特征,1 个属性标签。数据集中共包含 4 大类 38 种攻击。

3.1 数据预处理

每条入侵数据的 3 个字符型属性必须转化为数字类型才可以用 ESSFC 算法进行计算。每条入侵数据的 1 个属性标签为了方便识别也转化为数字类型。例如将入侵数据第 2 维属性可以转化为 $tcp=1,udp=2,icmp=3,\cdots$;第 3 维属性可转化为 $private=1,domain\_u=2,http=3,\cdots$;第 4 维属性可转化为 $SF=1,RSTR=2,REJ=3,\cdots$;第 42 维属性可转化为 $normal=1,back=2,ipsweep=3,\cdots$。又考虑到入侵数据多个属性使用的是不同度量单位,可能造成小数据被大数据淹没的状况,直接影响聚类分析的结果。因此在数据预处理阶段需对数据进行标准化。本论文采取对数标准化法。首先,计算出各列的最小值 min_colj 和最大值 max_colj,然后,用式(16)进行标准化:

$$data[j]=\frac{log\left(\frac{idenominator}{min\_colj+displacement}\right)+log(colj+dispalcement)}{log\left(\frac{idenominator}{max\_colj+displacement}\right)+log(max\_colj+dispalcement)} \tag{16}$$

其中:$idenominator=2$,$displacement=1.5$。

3.2 实验结果与分析

实验从 KDDDCUP99 数据集共选取样本 20 000 条,其中正常数据占 4 000 条,异常数据 16 000 条。异常数据包括 7 种攻击,分别为:mailbomb(4 000),ipsweep(600),back(600),portsweep(800),smurf(4 000),snmpgetattack(4 000),mscan(2 000)。

为了算法性能对比分析,首先进行基于模糊 C 均值聚类算法[5]实验。从标准化后的 20 000 条数据中随机选取 16 000 条数据(41 维,第 42 维是属性标签,不用于聚类,只用于最后的入侵检测指标——检测率和误警率的计算)作为生成分类器的训练数据,同时保证 8 类数据都有。测试数据为剩余的 4 000 条,同时也保证包含 8 类数据。初始化聚类中心为 8 个指定的样本。迭代计数为 1 000 次。阀值为 0.000 1。本文共实验 5 次,检测结果为:平均检测率为 58.93%。

然后进行 ESSFC 入侵检测算法实验。具体实验时选取的参数为:群体尺寸 p 为 10,目标的特征样本的类别 m=8,维数 n=41。从标准化以后的 20 000 条数据中选取 16 000 条数据(41 维,第 42 维是属性标签,只用于 ESSFC 算法的初始化和最后的入侵检测指标——检测率和误警率的计算)作为生成分类器的生成数据。其中选取标签数据为 6 000 条,未知标签数据为 10 000 条,同时保证 8 类数据都有。分别在 8 类中各选取一个样本作为 8 个初始聚类中心初始值。在计算目标函数的式(14)中用来保持未知标签数据模糊聚类差异和标签数据错分误差平衡的调整参数 α 为:10 000/6 000=1.667。

检测阶段,选取的测试数据为剩余的 4 000 条,同时也保证有 8 类数据。将 4 000 条测试数据都分类完成后,最后用检测阶段产生的结果和模拟数据集中相应数据的类标签,计算 ESSFC 入侵检测算法的检测指标。取 α=1.667,p=10,进化代数分别为 20,40,60,80,100,实验结果见表 1。由表 1 可知,平均检测率随着进化代数的增加而缓慢增加。

表 1 检测率与进化代数的关系

进化代数	20	40	60	80	100
检测率/%	80.65	81.41	81.49	82.35	82.66

改变生成 ESSFC 分类器中的生成参数 α。这次实验只改变这 16 000 条数据中标签数据和未知标签数据的比例。群体大小 p 为 10,特征样本的类别 m=8,维数 n=41。参数分别取 7,4,3,2,1.67。实验结果见表 2,由表 2 可知,平均检测率随着调整参数的减少而增加。

表 2 检测率与生成参数 α 的关系

调整参数	7	4	3	2	1.67
检测率/%	72.21	75.62	77.89	79.04	80.65

利用 KDDDCUP99 数据,通过进行基于模糊 C 均值聚类入侵检测算法和 ESSFC 入侵检测算法实验,实验结果表明 ESSFC 入侵检测算法比基于模糊 C 均值聚类入侵检测算法有高的检测率。

4 结论

本文针对入侵检测系统中,未知标签数据比标签数据较易获得,但只用未知标签数据去构造分类器的基于无监督式入侵检测算法(例如 C 均值聚类算法)效果较差情况,提出了一种基于进化半监督式模糊聚类入侵检测算法。该算法能够使用少量的标签数据和大量未知标签数据生成入侵检测系统分类器,可处理模糊类标签,不易陷入局部最优,适合于粗糙优化表面或拥有多个局部优化解时真值函数的优化。实验结果表明,该算法检测率高,适合并行结构的实现,相对于基于模糊 C 均值聚类的入侵检测算法有较高的检测率。

参考文献:

[1] 蒋建春,马恒太,任党恩.网络安全入侵检测:研究综述[J].软件学报,2000,11(11):1460-1466.

(下转 106 页)