一种适用于IDS的多次模糊迭代特征选择算法*)

李玲娟[1,2] 周桂芳[1] 王汝传[1,2]

(南京邮电大学计算机学院 南京 210003)[1] (苏州大学计算机科学与技术学院 苏州 215006)[2]

摘 要 本文针对入侵检测系统(IDS)被检测数据的特点,对适用于IDS的特征选择算法进行了研究,提出了一种基于分类的多次模糊迭代特征选择算法。该算法包括在属性空间中搜索特征子集、评估每个候选特征子集和分类这3个步骤,设计了与之相应的搜索算法和评估函数;算法通过多次迭代去除特征值集的冗余特征,得到精确度较高的特征值集;使用模糊逻辑得到与精确度要求相应的取值范围;由于单纯对数据进行操作,能比依赖于领域知识的算法更客观地分析数据。文内还对所提出的算法做了测试实验;并将实验结果与用可视化工具产生的特征可视化结果进行了比较。结果表明:该算法在IDS数据集上可取得良好的特征选择效果。

关键词 入侵检测系统,特征选择,模糊

A Multi-time Fuzzy Iterating Feature Selection Algorithm Adapting to IDS

LI Ling-Juan[1,2] ZHOU Gui-Fang[1] WANG Ru-Chuan[1,2]

(School of Computer, Nanjing University of Posts and Telecommunications, Nanjing 210003)[1]

(Computer Science & Technology School, Soochow University, Suzhou 215006)[2]

Abstract Based on the characteristics of detected data in IDS, feature selection algorithms adapting to IDS are studied in this paper, and a Multi-time Fuzzy Iterating Feature Selection Algorithm is proposed. This algorithm includes three steps, one is searching feature subsets from feature space, the other is valuating every candidate feature subset, and the last is classification. Corresponding search algorithm and valuation function are designed in the algorithm. The algorithm eliminates redundant features through multi-time iterating to get high precision feature value set, uses fuzzy logic to get the value range meeting the need of precision. This algorithm can analyze data more objectively than the algorithm with field knowledge for it only operates datasets. The paper also does some test experiments on the algorithm, and compares experiment results with feature visualization results from visualization tools. The results indicate: this algorithm can get good feature selection effect on IDS datasets.

Keywords IDS, Feature selection, Fuzzy

1 引言

入侵检测系统(IDS: Intrusion Detection System)是一种用来发现外部攻击和合法用户滥用特权的系统。它根据用户的历史行为,基于用户的当前行为,完成对入侵的检测,并留下证据,为数据恢复和事故处理提供依据。就数据搜集机制而言,IDS可以分为基于主机的和基于网络的两种类型。前者主要依赖于系统和应用程序的审核日志或者由可载入系统内核的模块分析提供数据源;后者的主要数据是捕捉到的网络数据包。就检测技术而言,IDS有误用检测、异常检测、混合检测。从概念上讲,这3种方式都离不开下列两个组件:特征和检测模型的建模算法。而判别特征的选择又是建立检测模型的基础,特征选择的效果对整个IDS的性能和检测准确率有着重要的影响。

特征选择的目的是找到最佳特征子集。然而它是经典的NP-hard问题[1],因而目前还没有高效的最佳特征选择算法。许多研究人员在对特征选择的研究中,将选择结果的标准定为特征数最少,或者是得到的规则最简,或选择量最大[2]。但是这些方法都没有考虑到不同领域数据的特点以及用户需求的灵活性。

数据挖掘技术是从已知数据集中挖掘出有用信息的技术,在商务分析、智能决策等领域都有着广泛的应用。将数据挖掘技术应用到IDS中,对有效地进行特征选择,建立合适的检测模型,最终提高IDS的入侵检测能力、降低其误报率和漏报率都有着十分重要的意义。

特征选择是数据挖掘在IDS中的主要应用之一。本文针对入侵检测系统被检测数据的特点,在基于数据挖掘分类方法的特征选择算法中引入模糊聚类方法,提出了一个结合分类和模糊聚类方法的适用于IDS的多次模糊迭代特征选择算法,用户可以根据实际需要更改阈值λ,以得到满意的特征选择结果。实验表明,该算法对IDS数据集具有较好的特征选择效果。

2 常用特征选择算法分析

特征选择的目的在于降低数据集的属性维数,剔除数据中的冗余成分和不相容性,提取关键信息。

*)本文受国家自然科学基金(60173037和70271050)、江苏省自然科学基金(BK2005146)、江苏省高技术研究计划(BG2004004)、江苏省计算机信息处理技术重点实验室基金(kjs050001)、江苏省高校自然科学研究计划(05KJB520092)资助。**李玲娟** 博士生,副教授,主要研究方向为数据挖掘、信息安全、网格计算等;**周桂芳** 硕士研究生,研究方向为数据挖掘、信息安全;**王汝传** 教授、博士生导师,研究方向为计算机软件理论、计算机网络及信息安全、移动代理技术等。

特征选择算法在特征子集空间中进行搜索，必须解决以下4个方面的问题：

1）搜索起始点（或搜索的方向）的选择。包括从空集开始的前向搜索、从全集开始的反向搜索、从中间某一点开始的双向搜索和从任意方向开始的随机搜索等。

2）搜索策略的确定。常用的搜索策略有完全搜索、启发性搜索、不确定性搜索这3种。其中，完全搜索为了不丢失最优解，通常会搜索到每一个属性子集；启发性搜索通过使用启发性信息避免了简单的完全搜索，速度快但得到的是近优解；不确定性搜索策略随机产生下一个待评价的子集，而不是顺序产生，不需要算法运行结束就可以得到解，但不知道什么时候能得到最优解，只知道下一个解会比现在的好。通常应该根据搜索空间的大小选择不同的搜索策略。

3）评价方法的选择。对于一个新产生的特征子集，必须依据某一标准对它进行评价，根据评价值决定下一步的搜索方向或停止搜索。评价的方法有很多，如：信息增益、一致性评价、正确性评价等。在搜索过程中评价方法也用来判断哪个属性应该加入或删除。

4）停止标准的确定。特征选择算法必须决定在什么时候停止对子集空间的搜索。这与具体的评价标准有关。一般情况下搜索算法会在无论添加或减少任何属性都只能得到更差的评价值的时候停止。也可事先规定算法执行的时间，规定随机搜索的次数或预设某一评价门限值等。

常用的特征选择算法有ABB算法、Relief算法和LVW算法，其中：

ABB(an automated branch and bound backward search algorithm)算法[3]采用后向搜索（深度优先）、一致性评价方法、完全搜索策略；它以找不到满足一致性要求的更小属性子集为停止标准。

Relief算法[4]专门处理涉及到的多个属性相互关联、相互作用的问题。许多归纳算法（如ID3和C4.5）在选择相关属性方面不很令人满意，而完全搜索的效率又太低。Relief方法基于对相关性的统计来选择属性。Relief算法能够处理离散的和连续的数据，但它对冗余属性的去除大多无能为力，而且一般只能用于二值类别数据。Kononenko扩展了原始Relief算法，使它能够处理有噪声的、不完整的和多类别的数据[5]。

LVW算法的搜索方向和搜索策略是随机的；评价方法属于正确性评价，就是用某一种机器学习算法或者分类器来评价属性子集；停止标准是预先指定的循环次数。文[6]中指出随机算法的好坏对结果是有影响的，LVW采用的是与时间相关的算法。

3 适用于IDS的多次模糊迭代特征选择算法

3.1 IDS数据源的特点

IDS中的数据源有其特点。以MIT林肯实验室于1998年第一次搜集和发布的用于评估网络入侵检测系统的数据集为例，Stolfo等根据领域经验分析总结了41个特征，这些特征主要分为4类：

(1)单个TCP连接的基本特征(9个)；

(2)根据领域知识所建议的一个连接内的内容特征(13个)；

(3)使用一个两秒窗口计算得到的流量特征(9个)；

(4)使用一个有100个连接窗口计算得到的基于主机的流量特征(10个)。

这其中很多是数值型的特征，出现攻击时，相应特征的值会发生异变，这些特征可用作判别特征，下文称之为关键特征。对于不同的攻击，关键特征不同，但每种攻击的关键特征都只有几个。因此，对IDS数据源的特征进行以选出关键特征为目标的优化选择，可以提高IDS处理速度和效率，而时效性和快速处理能力正是评价IDS整体性能的重要因素，特别对IDS这种数据集非常大的系统而言，减少一个次要特征所带来的性能提升是显而易见的。

3.2 适用于IDS的多次模糊迭代特征选择算法

针对IDS中数据源的特点，本文提出了采用多次模糊迭代的方法进行特征选择。概括地讲，就是首先使用预设定精度对训练数据集进行规则推理，得到规则集；然后对规则集进行统计，统计出对规则集贡献大的特征子集，即关键特征子集。

从上述过程可以看出，特征选择的关键在于要得到一个准确的分类器，下面给出分类器的数学描述：

设一个分类器系统由一个四元组$\langle U,F,J,R\rangle$构成，其中：$F=\{c_1,c_2,\cdots,c_n\}$表示数据集中的属性集合；$U=\{X\}$，$X=\{x_1,x_2,\cdots,x_n;c_i\}$表示数据实例；$J$为实例所属的类别集合，$c_i\in J$；$R$是分类映射函数。

分类器的一个重要性能参数是误分类信息损失，对于每个类$c_i\in J$，误分类信息损失$Err_J=\sum_{a\in J}Lc_jc_j$，此处$L_{CjCj'}$是将$c_j$类的实例误分成$c_{j'}$实例($j\neq j'$)时的损失。特征选择的目的就是建立一个新的分类系统$\langle U',F',J,R'\rangle$，$U'\in U$，$F'\in F$并且使得$ERR(R')\leqslant ERR(R)$。

目前带入侵检测领域知识的特征选择算法已经很多，本文提出的多次模糊迭代特征选择算法是一种不带领域知识的特征选择算法。图1是其选择过程。

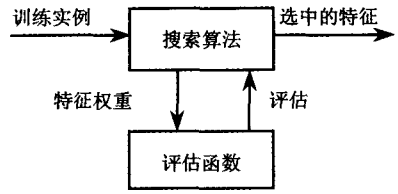


图1 特征选择过程

本文提出的算法用到以下定义：

定义1 设特征数据集$U=\{x_1,x_2,\cdots,x_n\}$，类别集合C_i($i=0,1,\cdots,n$)，则某一特征的加权平均值为

$$\bar{x}=\frac{\sum_{i=0}^{n}w_ix_i}{\sum_{i=0}^{n}w_i} \tag{1}$$

其中w_i为x_i出现的次数，n为数据集合的大小。

定义2 比较$|\bar{x}-xmin|$，$|xmax-\bar{x}|$值的大小，取两者之间最小值，并设为$radii$，则对特征数据集U进行剪枝，得到

$$U'=\{x_i/x_i\in(\bar{x}-radii,\bar{x}+radii)\} \tag{2}$$

定义3 设在类别集合C中存在C_i和C_j两类类别特征，则C_i赢得C_j概率为

$$Pro_{win}=\frac{\sum_{i=0}^{n}x_i\in U'_i\&x_i\in C_i\cap\sum_{i=0}^{n}x_i\in U'_j\&x_j\in C_j}{\sum_{i=0}^{n}x_i\in U'_i\&x_i\in C_i} \tag{3}$$

本算法采用的前向搜索策略，由空集S开始，逐个加入新的特征而组成特征子集。算法描述如下：

输入:测试数据 U_1

Pro_{uin}-可调整的赢得概率

初始化:$S=\{\}$

for ($i=0, i<n, i++$)

{

对每个特征数据集 U_i 进行剪枝得到 U'_i; 比较 $U'_i \in C_i$(C_i 为某种攻击类别)和 $U'_i \in normal$ 算得实际赢得概率 $Pro_{uin'}$;

if $Pro'_{uin} < Pro_{uin}$ then

$S+=Ui'$属于的特征 C_i;

}

输出:特征子集 S

4 算法测试及性能分析

本文对所提出的算法做了测试实验;并将实验结果与用可视化工具直接产生的特征可视化结果进行了比较。

4.1 实验数据及环境的选择

由于训练数据的多少以及训练数据之间的离散性会严重影响特征选择的精度。当训练数据之间的相似性太高时,会导致特征选择结果比较模糊,从而得到一个缺乏代表整个特征集合的特征子集;训练数据太大时,又会影响算法的执行时间。因此,本文的测试实验在训练数据的选择上,沿用了被广泛采用的训练集占整个数据集的 1/3 和测试集占整个数据集 2/3 的比例,所选训练数据来源于 3.1 节述及的 MIT 林肯实验室的入侵检测评估程序的数据集。它包括了网络数据包所能蕴涵的所有信息,这些记录是从带有模拟攻击的原始网络数据流中采集和处理得到的。每个记录由 42 个特征组成,最后一个属性标记攻击类型。数据集中包括的攻击类型有 22 种,分别属于以下 4 个范畴:

DOS(Denial of Service)——拒绝服务攻击,如 Back;

R2L——远程系统未授权访问,如 ftp_write;

U2R——未授权使用或非法使用超级用户权限,如 bufer_overflow;

Probing——监视或刺探系统脆弱性,如 ipsweep。

数据实例形式如下:

3058, udp, other, SF, 147, 105, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0.00, 0.00, 0.00, 0.00, 1.00, 0.00, 0.00, 255, 1, 0.00, 0.63, 0.88, 0.00, 0.00, 0.00, 0.00, 0.00, normal;

0, tcp, private, S0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 250, 12, 1.00,1.00, 0.00, 0.00, 0.05, 0.05, 0.00, 255, 12, 0.05, 0.06, 0.00, 0.00, 1.00, 1.00, 0.00, 0.00, neptune;

0, icmp, ecr_i, SF, 1032, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 511, 511, 000, 0.00, 0.00, 0.00, 1.00, 0. 00, 0. 00, 255, 255, 1. 00, 0. 00, 1. 00, 0. 00, 0.00, 0.00, 0.00, 0.00, smurf。

本文的算法测试实验环境为:PentiumIV2. 4、256M RAM,WindowsXP2. 0。

4.2 测试用的算法及测试结果

实验中设计的对多次模糊迭代特征选择算法结果进行测试的算法如下:

输入:测试数据 U_2

特征子集 S

初始化:$count_1=0$; $count_2=0$

for ($i=0$, $i<n$, $i++$)

{

用上述算法得到的关键特征对每一条测试数据判别所属类别,判为属于 C_i

$count_1++$; //对测试数据的实际类别 C_i 进行统计

}

for ($i=0, i<n, i++$)

{

用上述算法得到的关键特征对每一条测试数据判别所属类别,判为属于 C_i;

在数据集 U_2 中取得实际类别 C_j;

if $C_i==C_j$

then $count_2++$; //对测试数据的实际类别 C_j 和算得类别 C_i 比较

}

$$v_i=\frac{count_2}{count_1}*100\%;$$

输出:v_i

在实验中,当选择上述数据集的前 20000 条记录作为训练集,后 10000 条记录作为测试集时,选取阈值 $Pro_{win}=0.8$,迭代次数=2 时,算法执行时间为:120s,执行结果如图 2。

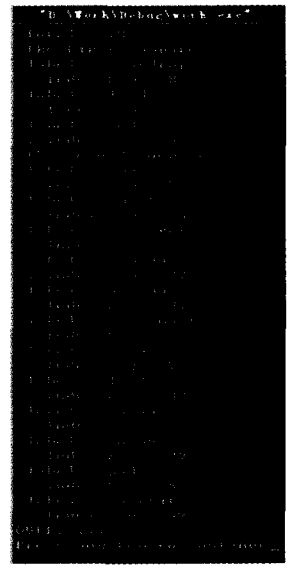
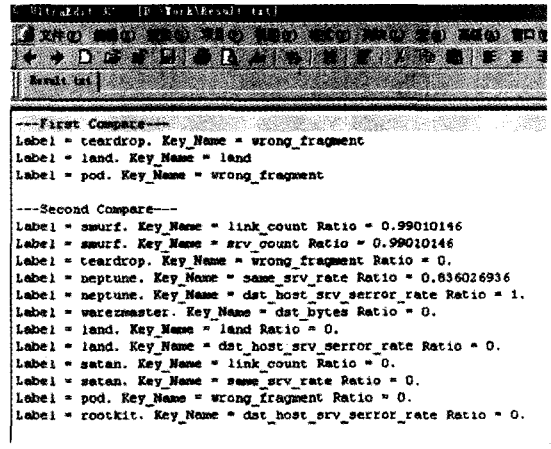


图 2 实验结果示例



```
---First Compare---
Label = teardrop. Key_Name = wrong_fragment
Label = land. Key_Name = land
Label = pod. Key_Name = wrong_fragment

---Second Compare---
Label = smurf. Key_Name = link_count Ratio = 0.99010146
Label = smurf. Key_Name = srv_count Ratio = 0.99010146
Label = teardrop. Key_Name = wrong_fragment Ratio = 0.
Label = neptune. Key_Name = same_srv_rate Ratio = 0.836026936
Label = neptune. Key_Name = dst_host_srv_serror_rate Ratio = 1.
Label = warezmaster. Key_Name = dst_bytes Ratio = 0.
Label = land. Key_Name = land Ratio = 0.
Label = land. Key_Name = dst_host_srv_serror_rate Ratio = 0.
Label = satan. Key_Name = link_count Ratio = 0.
Label = satan. Key_Name = same_srv_rate Ratio = 0.
Label = pod. Key_Name = wrong_fragment Ratio = 0.
Label = rootkit. Key_Name = dst_host_srv_serror_rate Ratio = 0.
```

图 3 result. txt 的显示结果

其中 Label 代表攻击名称，Index_Key 代表关键特征；执行结束时，测试算法自动产生各种攻击的参数文件。其中的一个重要文件是 result. txt 文件，该文件给出了多次模糊迭代特征选择算法在测试集上算得的正确率(见图 3)。

由图 3 可见，多次模糊迭代特征选择算法对关键特征的选择正确率在 83.6%以上，有的甚至达到 99%，而显示为 0 的数据表示在测试集中并未找到训练集中存在的攻击类型。当然这只是其中一种情形，当训练集和测试集的选择不同，以及 Pro_{win} 和迭代次数的选择不同时，算法结果、运行时间、算法精度都会发生改变。

4.3 用可视化方法验证算法的客观性和有效性

可视化已经成为目前信息处理系统必不可少的技术。在特征选择中可视化是一种事前选择方法。本文用可视化方法客观展示出的攻击与非攻击状态下主要特征的变化情况来验证所提出的特征选择算法的有效性。为了客观地观察入侵检测的特征，本文基于上述数据集的前 3000 条记录，给出了可视化的结果，并将之与 4.2 的测试结果做了比较。

图 4 展示的是将本文使用的数据源的前 3000 条记录作为样本导入可视化工具 ggobi 中，特征 land 值的取值可视化结果。该图 $x=2506$ 时对应的 land=1(见右上方的点)为异变点，而该记录对应的攻击类型就是 land 攻击，这与用本文提出的算法所挖掘的知识(见图 2)吻合。这也是第一次迭代的结果(也就是不加入阈值控制的结果)。

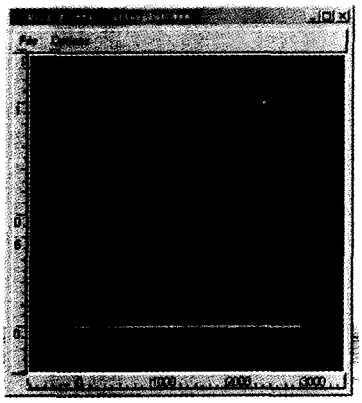


图 4 数据源前 3000 条记录中 land 特征取值的可视化显示

图 5 数据源前 3000 记录中 link_count 特征取值的可视化显示

图 5 展示的是对应于 smurf 攻击的可视化结果。在该数据源前 3000 条记录中，smurf 攻击对应特征 link_count 的取值在实际 excel 表中只存在于 3 个范围段，分别是 275～390，1381～1617，2933～3000。图 5 在对应的 $x \in (275,390) \cup (1381,1617) \cup (2933,3000)$ 时 link_count 出现异变，而这些记录对应的攻击类型就是 smurf 攻击，这也与用本文提出的算法所挖掘的知识(见图 2)非常吻合，这是第二次迭代的结果(也就是加入阈值控制的迭代结果)。

以上两例表明，本文提出的多次模糊迭代特征选择算法对各种攻击选出的关键特征与可视化的客观展示结果相符，由此可见算法的客观性和有效性。

结束语 本文针对 IDS 中被检测数据的特点，提出了一种多次模糊迭代特征选择算法，经理论分析和实验验证，该算法有以下特点：

1) 多次迭代。在去除特征值集合的冗余特征时，多次计算 *radii* 得到精确度很高的特征值集合。

2) 模糊逻辑。计算 Pro_{win} 时，采用模糊逻辑思想，可根据对精确度的要求得到取值范围。

3) 领域知识无关性。由于目前很多特征选择算法都加入了领域知识进行筛选，这样算法对领域知识的依赖性就增大，而本文提出的算法正好解决了这个问题，它单纯对数据进行操作无须领域知识，能更客观地分析数据。

参 考 文 献

1 Wong S K M, Ziarko W. On optional decision rules in decision tables [J]. Bulletin of Polish Academy of Science, 1985, 33: 693～696

2 常犁云，王国胤，吴渝. 一种基于 Rough Set 理论的属性约简及规则提取方法[J]. 软件学报，1999，10(11)：1206～1211

3 Liu Huan Scalable S R. Feature selection for large size database. In: Proceedings of the Fourth World Congress on Expert Systems. Morgan Kaufmann Publishers, 1998

4 Kira K, Rendell L. A practical approach to feature selection. In: Sleeman, Edwards P, eds. Proceedings of the Ninth International Conference on Maching Learning

5 Kononenko I. Estimating attributes : Analysis and extension of RELIEF. In: Proceedings of the European Conference on Machine Learning, 1994

6 Setiono, Liu Huan. Feature selection and classification--a probabilistic wrapper pproach. In: Proceedings of the Ninth International Conference on Industrial and Engineering Applications of AI and ES

7 Liu Huan, Setiono R. Chi2 Feature selection and discretization of numeric attributes. In: Proceedings of the Seventh IEEE International Conference on Trools with Artificial Intelligence, 1995

8 Fayyad U, Irani K. Multi-interval discretization of continuous-valued attributes for classification learning . In: Proceedings of the Thirteenth Internationaal Joint Conference on Artificial Intelligence, 1993

9 Hall M A. Correlation-based Feature Selection for Machine Learning: [PhD dissertation]. Department of Computer Science, University of Waikato, 1999

10 Zhao Jun, Wang Guo-Yin, Wu Zhong-Fu, et al. The study on technologies for feature selection. In: Machine Learning and Cybernetics, Proceedings, 2002 International Conference 2002

11 Amoro E G. Intrusion detection an Introduction to Internet Surveillance, Correlation, Traps, TraceBack, and Response. Http://www. intrusion. net, 1999

12 Rizzi A, Panella M, Mascioli F M F, et al. Automatic feature selection for adaptive resolution classifiers. In: Fuzzy Systems, 2002. FUZZ-IEEE'02, Proceedings of the 2002 IEEE International Conference on, Volume: 1, May 2002

13 Kummar S, Spafford H. A Pattern Matching Model for Misuse Intrusion Detection. In: Proceedings of 17th National Computer Security Conference, 1994. 11