入侵数据特征并行选择算法

于 泠[1,2]，陈 波[2]

(1. 解放军理工大学通信工程学院 南京 210007; 2. 南京师范大学计算机系 南京 210097)

【摘要】用知识的条件粗糙熵定义了特征的相对重要性，提出了一种基于条件粗糙熵的入侵数据特征并行选择算法。算法首先将入侵数据决策表划分成多个子表，然后利用特征的相对重要性对各子表并行求解，最后以子表选出的局部特征为基础求得原决策表的约简。实验表明，该算法适用于大规模的入侵数据集，选出的特征属性不仅可以大大减少数据在存储、分析以及各组件共享中的代价，还能够保持并提高入侵分类的准确性。

关 键 词 特征选择; 入侵检测; 并行算法; 粗糙熵; 粗糙集

中图分类号 TP393.08 文献标识码 A

Parallel Algorithm of Feature Reduction in Intrusion Data

YU Ling[1,2] and CHEN Bo[2]

(1. Institute of Communication Engineering, PLA University of Science and Technology Nanjing 210007;
2. Department of Computer Science, Nanjing Normal University Nanjing 210097)

Abstract This paper defines the importance of attack features using conditional rough entropy of knowledge and presentes a parallel algorithm of optimal feature selection in intrusion data based on conditional rough entropy. The algorithm divides the decision table of intrusion data into several sub-tables, and then the conditional rough entropy is used for the parallel computing of the sub-tables. Finally, the original decision table reduction is obtained based on the part reduction results from the sub-tables. The proposed algorithm has good performance and is good at dealing with the huge volume of data. The experimental results show that it is effective to reduce the storage requirements of the dataset and the computational cost, and it can increase the detection speed and without sacrificing the detection correctness by using the reduced feature subset.

Key words feature selection; intrusion detection; parallel algorithm; rough entropy; rough set

在入侵检测系统中，需要对大规模的网络数据流或主机审计信息进行复杂的数据分析(尤其是神经网络的构建和训练)，通常需要耗费大量的系统代价。入侵特征选择或提取技术正是用于从原有的庞大的入侵数据集中获得一个最优子集。

属性约简是粗糙集理论中的重要内容之一[1]。本文讨论的属性约简为入侵数据集中的最优特征选择，即从相关的特征集合中选取不含多余特征并保证分类正确的最小特征集。已证明求最小约简问题是NP-hard的，因此寻求快速、高效的约简算法已成为主要的研究课题。现有的属性约简算法包括基于差别矩阵及其改进的算法[2]、基于正区域[3]以及用信息熵作为选择重要属性的启发式属性约简算法[4-6]。

本文在上述粗糙集相关理论成果的基础上，用知识的条件粗糙熵定义入侵数据特征的相对重要性，设计了一种基于条件粗糙熵的入侵数据特征的并行选择算法。

1 算法理论基础

1.1 网络入侵数据决策表知识表达系统

本文研究的对象是网络入侵数据决策表系统，属于特殊的知识表达系统[1]，可表示为：

$$T=<U, R, V, f> \qquad R=C\cup D，D\neq\varnothing$$

式中 U是入侵数据样本集合；C为入侵数据特征(条件属性)集合；$D=\{d\}$为入侵类型(决策属性)集合；$V=\bigcup_{r\in R}V_r$ 为属性值的集合，V_r表示属性$r\in R$的属性值范围，即r的值域；f: $U\times R\rightarrow V$是一个信息函数，它指定U中每一个对象的属性值。

1.2 条件粗糙熵

下面给出入侵数据特征集(属性集合)的条件粗

收稿日期：2006 – 07 – 04；修回日期：2006 – 12 – 09
基金项目：江苏省普通高校自然科学研究计划(06KJD520101)；江苏省自然科学基金(BK2005135)
作者简介：于 泠(1971 –)，女，博士生，讲师，主要从事网络信息安全、数据挖掘、人工智能等方面的研究.

糙熵的定义以及相关定理。

定义 1　(条件粗糙熵)[1]　P为U上的一个条件属性子集合，$U/\mathrm{IND}(P)=\{X_1,X_2,\cdots,X_n\}$，$U/\mathrm{IND}(\{d\})=\{Y_1, Y_2,\cdots,Y_m\}$，则决策属性$\{d\}$相对于条件属性子集合$P$的条件粗糙熵为：

$$H(\{d\}\,|\,P)=-\sum_{i=1}^{n}\frac{|X_i|}{|U|}\sum_{j=1}^{m}\frac{|Y_j\cap X_i|}{|X_i|}\log\frac{|Y_j\cap X_i|}{|X_i|}$$

根据以上定义，给出定理 1。限于篇幅，定理均未列出证明。

定理 1　设$T=<U, R, V, f>$是一个入侵数据决策表系统，$P_1,P_2\subseteq R$，且$U/\mathrm{IND}(P_1)=\{X_1,X_2,\cdots X_n\}$；$U/\mathrm{IND}(P_2)=\{X_1, X_2,\cdots, X_{i-1}, X_{i+1},\cdots, X_{j-1}, X_{j+1},\cdots, X_n, X_i\cup X_j\}$，$U/\mathrm{IND}(\{d\})=\{Y_1,Y_2,\cdots,Y_m\}$，其中$U/\mathrm{IND}(P_2)$是将$U/\mathrm{IND}(P_1)$中的任意两个等价块$X_i$和$X_j$合并为$X_i\cup X_j$而得到的，则$H(\{d\}|P_2)\geqslant H(\{d\}|P_1)$。

该定理说明，如果将属性集的分类进行合并，将可能导致条件粗糙熵的增加。换句话说，如果一个属性不能为属性子集的分类增加任何信息，就可以将这个属性约简。

由定理1可以得到如下推论。

推论　设$T=<U, R, V, f>$是一个入侵数据决策表系统；A为C经过特征选择后得到的特征集合，C_0是核。如果$a_i\in A\backslash C_0$是任意一个不能被约简的特征属性，则有：

$$H(\{d\}\,|\,C_0)>H(\{d\}\,|\,C_0\cup\{a_1\})>\cdots>$$
$$H(\{d\}\,|\,C_0\cup\{a_1\}\cup\{a_2\}\cup\cdots\cup\{a_i\}\cup\cdots)>\cdots>H(\{d\}\,|\,A)$$

该推论说明，如果特征选择以核为起点，那么在选择过程中，向特征子集C'中不断添加不能约简的特征a得到的条件粗糙熵$H(\{d\}|C'\cup\{a\})$是单调递减的。该推论给出了入侵数据特征选择的方法，即以核作为初始状态，通过不断增加特征来得到最终选择结果。

定理 2[1]　设$T=<U, R, V, f>$是一个入侵数据决策表系统，U是在C上相对于$\{d\}$是一致的，则对于属性$r\in C$，r是C相对于决策属性集合$\{d\}$不必要的，其充要条件是：$H(\{d\}|C)=H(\{d\}|C\backslash\{r\})$。

定理2给出了特征选择方法的终止条件。

1.3　基于条件粗糙熵的入侵特征重要性

定义 2　(入侵特征的相对重要性)　设$T=<U, R, V, f>$是一个入侵数据决策表系统，特征$a\in C$在C中对$\{d\}$的重要性定义为：

$$\mathrm{SGF}(a, C,\{d\})=H(\{d\}|C\backslash\{a\})-H(\{d\}|C)$$

定义2说明，在已知C的条件下，$\mathrm{SGF}(a, C,\{d\})$的值越大，特征a对于决策$\{d\}$就越重要。

定义 3　(相对核)　设$T=<U, R, V, f>$是一个入侵数据决策表系统，C中所有对$\{d\}$是必要的特征组成的集合称为特征集合C相对于$\{d\}$的相对核，记作$\mathrm{CORE}(C,\{d\})$。

性质 1　$0\leqslant\mathrm{SGF}(a,C,\{d\})\leqslant 1$。

性质 2　在已知C的条件下，当且仅当$\mathrm{SGF}(a,C,\{d\})>0$，入侵特征$a$是必要的。

性质 3　$\mathrm{CORE}(C,\{d\})=\{a\in C|\mathrm{SGF}(a,C,\{d\})>0\}$。

定义 4　(入侵特征的选择)　设$T=<U, R, V, f>$是一个入侵数据决策表系统，$A\subseteq C$。如果任一特征b不能为特征子集A的分类增加任何的信息，即有$H(\{d\}|A\cup\{b\})=H(\{d\}|A)$，则$A$为$C$的相对于$\{d\}$的特征选择。

1.4　数据划分

并行算法会涉及到数据划分以及对分解的数据并行操作的问题。本文采用的划分方法是：将入侵特征决策表均等地划分为k个子表，每个子表具有$|U|/k$条数据记录，共需k个从进程分别对k个子表进行局部特征选择。为了使局部选择的结果更具可信性，每个子表中各种入侵类型所占比例与原始表保持一致。

1.5　全局特征选择

并行特征选择还应考虑如何根据局部选择结果得到全局选择的结果。为描述清楚，本文做如下约定：CORE_i表示第i个从进程局部选择出的核；$\mathrm{CORE}_{\mathrm{global}}$表示全局的核，根据粗糙集理论，扩大考察范围不会减少核属性，全局核必然包含所有的局部核[7]，即$\mathrm{CORE}_{\mathrm{global}}=\bigcup_{i=1}^{k}\mathrm{CORE}_i$。$\mathrm{RED}_i$表示第$i$个从进程局部特征选择的结果(含核)；$\mathrm{FEA}_{\mathrm{com}}$表示除核外各局部特征选择结果中的共同特征集，即$\mathrm{FEA}_{\mathrm{com}}=\bigcap_{i=1}^{k}(\mathrm{RED}_i-\mathrm{CORE}_i)$，该部分特征具有较大的分辨能力，是全局选择中不可缺少的；$\mathrm{FEA}_{\mathrm{can}}$表示各局部特征选择结果中除核和共同特征外的其他特征，称为候选特征集，即$\mathrm{FEA}_{\mathrm{can}}=\bigcup_{i=1}^{k}(\mathrm{RED}_i-\mathrm{CORE}_i-\mathrm{FEA}_{\mathrm{com}})$。$\mathrm{RED}_{\mathrm{global}}$表示全局特征选择结果(含核)，即$\mathrm{RED}_{\mathrm{global}}=\mathrm{CORE}_{\mathrm{global}}\cup\mathrm{FEA}_{\mathrm{com}}\cup\mathrm{Part_of}\,(\mathrm{FEA}_{\mathrm{can}})$，其中$\mathrm{Part_of}\,(\mathrm{FEA}_{\mathrm{can}})$表示从$\mathrm{FEA}_{\mathrm{can}}$中选择出的特征集。选择方法是：从各子表中抽取$|U|/k^2$条入侵数据记录组成第$k$+1张子表，以$\mathrm{CORE}_{\mathrm{global}}\cup\mathrm{FEA}_{\mathrm{com}}$作为核，根据入侵特征相对重要性从$\mathrm{FEA}_{\mathrm{can}}$中选取特征，得到最终全局结果。

2 特征选择并行算法

根据上述理论，本文提出的入侵数据特征的并行选择算法描述如下。

2.1 主进程(p_{master})处理算法

1) 输入入侵数据决策表；

2) 将表平均分成k个子表；

3) 将各子表Sub T_i发送给相应的从进程p_i；

4) 对k个子表执行以下操作：(1) 接受各子表发回的局部核和局部选择结果；(2) 计算全局核$CORE_{global}$、全局共同特征集FEA_{com} 以及候选特征集FEA_{can}。

5) 从各子表中抽取$|U|/k^2$条入侵数据记录组成第k+1张子表；

6) 根据入侵特征的相对重要性，执行函数reduction (Sub_T_{k+1},$CORE_{global} \cup FEA_{com}$, FEA_{can})；从FEA_{can}中选取特征，得到最终全局选择结果。

2.2 从进程($p_1,p_2,\ldots,p_k$)处理算法

(1) 接受从主进程p_{master}发送来的子表Sub_T_i；

(2) 根据入侵特征相对的重要性，执行函数reduction(Sub_T_i, $\varnothing$,C)；从入侵数据特征集合C中选取特征，得到局部核$CORE_i$和局部选择结果RED_i，并发送回主进程p_{master}。

2.3 reduction()函数算法

reduction(T,$CORE_{init}$,FEA_{init})函数的功能是根据入侵数据决策表T，利用入侵特征相的对重要性从特征集FEA_{init}中得到最优特征子集。$CORE_{init}$表示核的初始状态。该函数算法伪代码如下：

```
CORE(C,{d})= CORE_init；C=FEA_init
//计算入侵数据决策表系统中的条件粗糙熵
Comput  H({d}|C)
if  CORE(C,{d})={}
{ //计算每个攻击特征a∈C的相对重要性
  for every a∈C
  {  SGF(a, C,{d})=H({d}|C\{a})–H({d}|C)
     if  SGF(a,C,{d})>0   //求相对核
         CORE(C,{d})= CORE(C,{d})∪{a}
  }
}
A=C\ CORE(C,{d}); B= CORE(C,{d})
if |B|!=0
{  Comput  H({d}| B)
   if H({d}| B)==H({d}| C)  Halt
}
flag=true
while flag
{  min=－1
  for every a∈A
  {  Comput  H({d}|B∪{a})
     //选择使条件粗糙熵最小的特征b
     if  H({d}|B∪{a})<=min
       { min= H({d}|B∪{a}); b=a}
  }
  A=A\b; B=B∪{b}
  if H({d}| B)==H({d}| C)
     {flag=false; Halt}
}
CORE_i= CORE(C,{d}); RED_i=B
```

3 特征选择结果分析及相关工作比较

3.1 实验方法与结果

本文所使用的实验数据是Kdd99。该数据集中每条连接记录包含41个特征及一个区分正常或异常的标记。41个特征属性分离散型和连续型两种，被划分到基本连接的特征、基于内容的特征、基于时间的流量特征和基于主机的流量特征4个不同的特征子集中。

下面通过实例来说明特征选择并行算法的执行过程。为了能清晰地描述算法执行过程，仅选择5个入侵特征C={Service(c_1)、Flag(c_2)、Num_failed_logins(c_3)、Su_attempted(c_4)和Root_shell (c_5)}，以及1个决策属性D和90条数据集U，如表1所示。

表1 一个训练数据集

U	C					D
No.	c_1	c_2	c_3	c_4	c_5	
1	0	0	3	0	0	0
2	0	0	0	1	0	0
3	1	0	0	1	0	0
4	1	0	2	0	0	1
5	0	0	4	0	0	1
6	1	0	1	2	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮
90	1	0	0	0	1	0

算法执行步骤如下：

(1) 主进程将90条数据平分成3个子表，依据本文数据划分的方法进行划分，并将相应子表传送给3个从进程；

(2) 从进程p_1～p_3的局部选择结果为$CORE_1$=

$\{c_1\}$, $RED_1=\{c_1,c_2,c_3\}$; $CORE_2=\{c_1\}$, $RED_2=\{c_1, c_2, c_3, c_4\}$; $CORE_3=\{c_1,c_3\}$, $RED_3=\{c_1,c_2,c_3,c_5\}$;

(3) 主进程根据各从进程局部选择结果进行全局选择，$CORE_{global}=\{c_1,c_3\}$，$FEA_{com}=\{c_2\}$，$FEA_{can}=\{c_4, c_5\}$，执行reduction(Sub_$T_{k+1}$, $CORE_{global} \cup FEA_{com}$, FEA_{can})函数，从FEA_{can}中选择出c_5特征，得到最终全局选择结果$RED_{global}=\{c_1,c_2,c_3,c_5\}$。

在本文的实验中，从kddcup.data _10_ percent.gz A10 % subset (2.1MB；75MB未压缩) 数据集中选取了37 500条记录作为实验基本数据，共分成15个子表。尽管Kdd99数据集中提供的数据已进行过预处理，但在使用的过程中还需要对其进行数据数值化、离散化处理。本文选用有监督算法Naive Scaler[1]。对经过处理后的数据，利用MPI的c++环境执行并行算法，选出了28个入侵特征(限于篇幅未列出)。

3.2 结果分析与相关工作比较

(1) 算法有效性分析。采用文献[8]提出的“数据浓缩”的两个指标对本文算法的实验结果进行分析，特征的蒸发率$E_{特征}$=(1–选择出的特征个数/原数据集特征个数)×100%=31.71%。数据蒸发率$E_{数据}$=(1–约简后的数据集数据量/原数据集数据总量)×100%=86.96%。根据文献[8]的经验值，$E_{特征}$>30%和$E_{数据}$>85%是令人满意的。

(2) 算法效率分析。由于寻找最小知识相对约简是一个NP-hard问题，其复杂性主要是由信息系统中的属性组合和数据量引起的[9]。令$|C|=m$，$|U|=n$，则本文算法的各从进程计算核CORE(C,{d})共需要计算m次SGF(a,C,{d})。计算一次划分的时间复杂度为$O((n/k)^2)$，所以计算一次SGF(a,C,{d})的时间复杂度为$O((n/k)^2)$。计算约简中，需要计算$H(\{d\}|B \cup \{a\})$的次数最多为$m+(m-1)+\cdots+1=m\times(m+1)/2=O(m^2)$，所以整个算法的时间复杂度为$T_1=O(m^2(n/k)^2)$。而串行化的算法时间复杂度为$T_2=O(m^2n^2)$。因此本文提出的算法更适合于大规模入侵数据集的特征选择。

在算法的空间复杂度方面，对比文献[4-5]中给出的基于分辨矩阵的约简算法，其时间和空间复杂度随着决策表的大小成指数变化，而且要求生成中间环节分辨矩阵。而本算法由于只需计算条件粗糙熵及入侵特征的相对重要性SGF，因此对存储量的需求远小于基于分辨矩阵的约简算法。

(3) 对检测性能的影响。分别用约简前后的数据集对文献[10]设计实现的基于小波神经网络的入侵检测系统进行测试，两种网络的拓扑分别是41-26-5(输入节点41个、隐层节点26个、输出节点5个)和28-16-5，训练过程耗时分别为306 s和198 s，所得到的ROC曲线(限于篇幅，略)表明在虚警率为10%的情况下，检测率均达到了90%以上，说明特征的约简并不会影响网络的分类性能，且可以缩短网络训练的时间。

4 结 束 语

实验表明，本文建立的基于条件粗糙熵的入侵数据特征并行选择算法是有效的，选出的特征集是不含多余特征并保证分类正确的最小特征集，达到了入侵知识表达空间的约简。本文的工作不仅可以减少系统存储、分析、共享的代价，降低构建神经网络系统的复杂性，简化训练集，减少检测时间，而且能够保持并提高入侵分类的准确性。

参 考 文 献

[1] 王国胤. Rough集理论与知识获取[M]. 西安: 西安交通大学出版社, 2001.

[2] WANG J , WANG J. Reduction algorithms based on discernibility matrix: the ordered attributes method[J]. Journal of Computer Science and Technology, 2001, 16(6): 489-504.

[3] 刘少辉, 盛秋戬, 吴　斌, 等. Rough集高效算法研究[J]. 计算机学报, 2003, 26(5): 524-529.

[4] 苗夺谦, 胡桂荣. 知识约简的一种启发式算法[J]. 计算机研究与发展, 1999, 36(6): 681-684.

[5] 刘启和, 李　凡, 闵　帆, 等. 一种基于新的条件信息熵的高效知识约简算法[J]. 控制与决策, 2005, 20(8): 878-882.

[6] 刘振华, 刘三阳, 王　珏. 基于信息量的一种属性约简算法[J]. 西安电子科技大学学报, 2003, 30(6): 835-838.

[7] 刘　山. 基于分治的属性约简复杂性分析[J]. 计算机工程与应用, 2004, 40(20): 102-103.

[8] 王　珏, 王　任, 苗夺谦, 等. 基于Rough Set理论的数据浓缩[J]. 计算机学报, 1998, 21(5): 393-400.

[9] GUAN J, BELL D. Rough computational methods for information systems[J]. Artificial Intelligence, 1998, 105: 77-103.

[10] 陈　波, 于　泠. 基于小波神经网络的服务器预警系统. 电子科技大学学报, 2005, 34(3): 343-346.

编　辑　熊思亮