维普资讯 http://www.cqvip.com

面向入侵检测的基于多目标遗传算法的特征选择*)

俞 研 黄 皓

(南京大学计算机科学与技术系,计算机软件新技术国家重点实验室 南京 210093)

摘 要 针对刻画网络行为的特征集中存在着不相关或冗余特征,从而导致入侵检测性能下降的问题,本文提出了一种基于多目标遗传算法的特征选择方法,将入侵检测中的特征选择问题视为多目标优化问题来处理。实验结果表明,该方法能够实现检测精度与检测算法复杂性的均衡优化,在显著提高检测算法效率的同时,检测精度也有所提高。

关键词 入侵检测,特征选择,多目标优化,遗传算法

Feature Selection Using Multi-Objective Genetic Algorithms for Intrusion Detection

YU Yan HUANG Hao

(Department of Computer Science and Technology, Nanjing University, Nanjing 210093)

(National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210093)

Abstract A feature selection method using multi-objective genetic algorithms is proposed to solve the problem of performance degradation of the intrusion detection, which results from the existence of irrelevant or redundant features among the feature set representing the network behavior. The method views the feature selection for intrusion detection as multi-objective optimization problem. The experimental results manifest that the best detection accuracy/complexity trade-off can be achieved. The detection accuracy is better, while the detection algorithm efficiency is improved remarkably.

Keywords Intrusion detection, Feature selection, Multi-objective optimization, Genetic algorithms

1 引言

在异常入侵检测中,系统需要处理海量的网络数据,而网络通信量的行为是由特征来刻画的。若在特征集合中存在大量冗余或不相关的特征,不但会在基于机器学习的入侵检测过程中产生噪声,使之难以区分异常的网络行为,检测精度下降,而且大量特征的存在会增加学习算法的搜索空间,导致算法的检测效率降低。因此,选择和提取能准确刻画网络行为的特征对于异常入侵检测系统就显得至关重要。

特征选择是根据给定的准则从一组特征中挑选出一些最有效的特征以降低特征空间维数,无用的、冗余的以及最少使用的特征将被从特征集合中删除[1]。在异常入侵检测中,特征选择的主要目标是在保证一定的检测精度的前提下,尽可能地减少用于分类的特征数量,以提高检测算法的效率。

传统的面向入侵检测的特征选择方法多是从单个特征的分类能力来考虑,选择包含最多分类信息的特征,而没有充分考虑特征数量与分类精度之间的平衡优化问题,即检测精度与算法复杂性的平衡优化问题。而采用多目标优化方法则可以同时兼顾检测精度与算法复杂性的优化,因而适用于面向入侵检测的特征选择问题。

在多目标优化问题中,通常采用的方法是将多个目标转换为单个目标,如线性加权和法等,将多个目标函数通过线性组合,组成单一的组合目标函数。然而采用线性加权和等方法,需要为每个目标函数分配适当的权值,最终解依赖于权值矢量的分配,权值矢量的设置是否合理将会严重影响到最终解的优劣。基于 Pareto 的多目标优化方法克服了加权和等方法中权值矢量的影响,从生成的 Pareto 最优前沿中选择最好的解作为最终解。另外,遗传算法所具有的并行搜索解空间以及较强的全局搜索能力,使其可以改善基于 Pareto 的多目标优化方法的性能。

因此,本文在分析从多目标优化的角度来解决入侵检测中的特征选择问题的基础上,提出了一种基于多目标遗传算法的入侵检测特征子集选择方法,利用有效的多目标遗传算法 NSGA-II 搜索 Pareto 最优前沿,作为特征选择的候选解集。同时,为了克服 Pareto 最优前沿中可能存在的泛化精度不高的问题,利用交叉验证方法选择具有最好的泛化精度的解作为构建异常入侵检测算法的特征子集。另外,由于支持向量机(SVM)具有快速分类的能力与良好的扩展性,以及其对入侵检测的适用性,选择 SVM 作为算法的适应度评价函数。试验结果表明,算法不但降低了入侵检测的复杂性,而且在一定程度上提高了检测精度。

本文的组织如下:第 2 节首先介绍了与入侵检测中特征选择算法以及多目标优化方法相关的研究工作;第 3 节给出了特征选择与多目标优化的概念与方法,并对其在入侵检测中的应用进行了分析;第 4 节则对面向入侵检测的基于多目标遗传算法的特征选择方法进行了描述;第 5 节为实验方法与结果,并对结果进行了分析;最后是全文的总结。

2 相关工作

由于入侵检测系统需要快速处理大量的网络数据,数据

*)国家 863 计划(2003AA142010);江苏省高技术计划(BG2004030)。俞 研 博士研究生,主要研究方向为网络安全、数据挖掘等;黄 皓 教授,博士生导师,主要研究方向为计算机软件、信息安全等。

维普资讯 http://www.cqvip.com

中蕴含着大量的特征，而特征之间存在着复杂的相互关系，因此，不相关的或包含噪声特征的存在使得检测变得愈发困难，阻碍了检测算法对数据的处理[2]。这就要求入侵检测系统必须能够消除不相关或噪声特征的影响，以提高检测的精度和效率，即入侵检测面临着特征选择问题。Sung[3]采用了支持向量机与神经网络方法，根据特征的重要性对其进行排序，试验结果显示网络数据中存在着大量并不重要的特征，删除这些特征并不会显著降低检测精度；Chebrolu[4]利用马尔可夫毯与判定树对特征进行选择，并使用约简的数据集分别构造基于贝叶斯网络与分类回归树的入侵检测模型，提高了检测的精度与效率。

然而，这些特征选择方法均是根据单个特征在入侵检测中的重要性进行特征选择，没有充分考虑到特征之间的相互关系对检测结果的影响，以及检测精度与特征数量之间的平衡问题。因此我们认为，在入侵检测中，可以将选择最优特征子集的问题，视为最大化检测精度与最小化特征数目的多目标优化问题来考虑。

多目标优化是研究在一定约束条件下的多个目标函数的极值化问题。在入侵检测中，检测精度与算法复杂性(或特征数量)之间通常存在着冲突，追求某一项最优可能导致另外的目标变差。同时，最优特征子集的选择并不唯一，因此适合于采用多目标优化方法来求解。另外，由于遗传算法在大的搜索空间中具有快速全局搜索的能力，因而被广泛应用于多目标优化问题[5,6]。

针对入侵检测中特征选择方法存在的不足以及对多目标优化的分析，本文考虑采用一种多目标遗传算法搜索 Pareto 最优前沿，并利用交叉验证方法从候选解集中选择最优解作为最优的特征子集。算法既解决了入侵检测中检测精度/算法复杂性均衡的多目标优化问题，又能够利用单独的验证集从候选解集中选择泛化精度最好的特征子集。

3 基本概念

3.1 特征选择

在网络中，大量的网络数据快速到达，这就要求入侵检测系统能够对入侵行为进行快速、准确的检测和响应。而在入侵检测算法中，描述网络数据的特征是否能够准确地刻画网络真实行为，决定了检测算法能否获得对正常行为与攻击进行分类所需的必要信息，检测精度也会受到影响。另外，特征的数量也确定了学习算法的搜索空间，不相关的、冗余的特征集合增加了算法的搜索空间，从而导致检测算法的学习效率不高。因此，从原始特征集合中选择最优的特征子集也就成为入侵检测必须面对的问题之一。

特征选择的任务就是根据给定的评价函数 J，从描述网络通信量行为的原始特征集 S 中(集合 S 的基数为 D)选择出最优特征子集 X，其中 X 的基数为 d，且 $d\leqslant D$。特征选择可以形式化表示为

$$J(X)=\min_{Y\subseteq S,\text{且}|Y|\leqslant D} J(Y) \tag{1}$$

其中：S 表示原始特征集合；X 表示选择的最优特征子集，$X\subseteq S$；$J(\cdot)$表示特征选择评价函数，不失一般性，此处采用了最小化准则函数。

由式(1)可以看出，为了提高入侵检测算法的性能，需要获得尽可能高的检测精度与尽可能小的特征子集，以提高算法的效率。显然，若选择的特征子集太小，会使得算法丢失一些具有较多分类信息的特征，因此目标之间存在着一定的冲突。而多目标优化问题就是在特征集合中选择一个最优的特征子集，使之满足各种约束的限制，又能实现多目标之间的均衡。因此，本文采用了多目标优化方法来解决入侵检测中的特征选择问题，并在下一节对基于遗传算法的多目标优化问题进行分析。

另外，特征选择方法根据其是否独立于学习算法可以分为两类：独立于学习算法的特征选择方法称为筛选方法，否则称为卷积方法。筛选方法独立于学习算法，因而其速度较快；卷积方法虽然速度相对较慢，但由于结合了学习算法，在特定的应用领域特征选择的效果好于筛选方法。由于在入侵检测中，特征选择的过程在构建入侵检测模型之前独立实现，对速度的要求相对较低，因此本文的特征选择方法采用了卷积方法。

3.2 基于遗传算法的多目标优化

多目标优化问题就是在一定的约束条件下，求解多个目标函数的最优化问题，可以形式化表示为

$$\min f(x)=[f_1(x),\ f_2(x),\cdots,\ f_p(x)]^T; \tag{2}$$
$$\text{s. t. } x\in X$$

其中 $x=(x_1,\cdots,\ x_n)^T$ 称为决策向量，X 为决策空间，$f_1(x)$，$\cdots$，$f_p(x)$称为目标函数，p 维向量$[f_1(x),\ f_2(x),\cdots,\ f_p(x)]^T$ 所在的空间称为目标空间。

在多目标优化问题中，通常多个目标之间存在着冲突，解不是唯一的，而是存在一个最优解集。集合中的元素称为 Pareto 最优解或非劣最优解。

定义 1(优劣性) 在最小化问题中，称目标向量解 $u=(u_1,\ u_2,\cdots,\ u_n)$优于 $v=(v_1,\ v_2,\cdots,\ v_n)$，当且仅当 u 部分小于 v，即$\forall i\in\{1,\ 2,\cdots,\ n\}, u_i\leqslant v_i \wedge \exists i\in\{1,\ 2,\cdots,\ n\}, u_i<v_i$。

定义 2(Pareto 最优解) 决策变量 $x_u\in X$ 称为 Pareto 最优解，当且仅当不存在决策变量 $x_v\in X$，使得相应的目标向量 $v=f(x_v)=(v_1,\ v_2,\cdots,\ v_n)$优于 $u=f(x_u)=(u_1,\ u_2,\cdots,\ u_n)$。

由所有 Pareto 最优解组成的集合构成了多目标优化问题的最优解集，也称为有效解集；Pareto 最优解相对应的目标向量称为非支配的(non-dominated)，由所有非支配的目标向量构成多目标问题的 Pareto 最优前沿。

传统的基于单目标优化的方法严重依赖于权值的选择或其他先验知识，并且至多只能获得单个 Pareto 最优解。而遗传算法具有较强的全局搜索能力，多个个体可以并行搜索解空间，并在一次运行中得到多个 Pareto 最优解，因而适用于多目标优化的情况。同时，采用基于 Pareto 的遗传算法[6]直接利用了 Pareto 支配关系来确定群体中个体选择的概率。该方法为群体中所有的非支配个体分配相同的选择概率，但是由于在选择过程中的随机误差，尽管 Pareto 最优解集中存在多个解，仍然会在群体数量有限时收敛到只有一个最优解。为此，Goldberg 提出了基于适应度共享机制的小生境方法，即通过反映个体之间的相似程度的共享函数来调解群体中各个个体的适应度，从而在之后的群体进化过程中，算法能够依据调整后的新适应度来进行选择运算，以维持群体的多样性，创造出小生境的进化环境，从而避免了“过早收敛”问题。

非支配排序遗传算法 NSGA-II[7]是一种带有精英策略的、快速的基于 Pareto 方法的多目标优化遗传算法，采用对多目标群体进行逐层分类的方法，每代选择配对之前先按照个体的非劣性进行排序，为群体中所有的非支配个体分配相

维普资讯 http://www.cqvip.com

同的选择概率，并引进了基于决策向量空间的共享函数法，具有 Pareto 最优解均匀分布，并允许存在多个不同等效解等优点。

通过以上对入侵检测中特征选择以及多目标优化等问题的分析，我们认为，采用 NSGA-II 算法生成的 Pareto 最优解集适合于作为入侵检测中最优特征子集的候选集合。

4 算法

4.1 算法描述

为了采用基于多目标优化遗传算法生成最优的入侵检测特征子集，首先需要对特征进行编码。而在入侵检测中，采用二进制编码方式是很自然的选择。群体中的每个个体均代表了特征选择问题的一个候选解，表示为一个 m 维的二进制向量。若某位的值为 1，则表示选择相应的特征，否则丢弃该特征。其中 m 表示初始特征集合中的特征数目。

根据前面的分析，可以知道，在入侵检测中的特征选择问题涉及到对两个目标的优化：最大化检测精度与最小化特征集合中的特征数目。其中特征数目可以通过计算候选解中位为 1 的个数来获得，因此我们主要考虑检测精度的计算。

由于特征选择中的卷积方法采用学习算法的结果来评价特征子集的优劣，故具有更好的准确性。因此，我们在算法中采用了特征选择的卷积方法。而在卷积方法中，群体中每一个代表候选特征子集的个体均需要学习算法对其进行评价。考虑到 SVM 具有较快的速度与良好的扩展性，已经被 Mukkamala 证明适合于入侵检测[8]，因此采用 SVM 作为分类器来评价候选特征子集可以快速地获得较好的入侵检测结果。同时，为了综合考虑算法的检测性能，我们采用了查准率(Precision)与查全率(Recall)的调和平均 F-value 值[9]作为检测精度的目标函数：

$$\text{Precesion}=\frac{TP}{TP+FP}$$

$$\text{Recall}=\frac{TP}{TP+FN}$$

$$F\text{-value}=\frac{2\cdot\text{Precision}\cdot\text{Recall}}{\text{Precision}+\text{Recall}} \tag{3}$$

其中：TP 表示正确检测到的入侵数据样本数；FP 表示被误报为入侵的正常数据样本数，即误报的样本数；FN 表示被检测为正常的入侵数据样本数，即漏报的样本数。

在本文中，我们采用最小化目标函数的方法来解决多目标优化问题，需将最大化检测精度的问题转化为求最小值的问题。因此，特征选择的多目标优化问题表示为

$$\min f(x)=[f_1(x),\ f_2(x)]^T;$$

$$\text{s.t. } x\in X$$

其中，$f_1(x)=1-\text{Fvalue}$；$f_2(x)=|F|$，F 为所选择的特征子集，$|F|$ 为子集的基数；X 为入侵检测的原始特征空间。

采用基于 SVM 的检测算法生成的结果作为目标的评价函数，需对每一个个体进行计算，必然会导致计算量的增加，因此我们为算法增加了记忆功能(bookkeeping)。由于在对个体(即特征子集)进行评价时，往往会发生对相同个体的重复计算，因此保存以往个体的计算结果，会显著缩短算法学习所需的时间，改善算法的执行效率。

在获得 Pareto 最优解集之后，需要根据入侵检测的具体要求，从解集中选择最优解，即最适合的特征子集。为了减少测试集的影响，提高检测算法泛化精度，我们选择了单独的验证集，利用交叉验证的方法来评价解集中的个体，从中选择最优的特征子集。

4.2 面向入侵检测的基于多目标遗传算法的特征选择算法

面向入侵检测的基于多目标遗传算法的特征选择算法如图 1 所示。

算法：面向入侵检测的基于多目标遗传算法的特征选择算法

输入：训练集 S_{train}，测试集 S_{test}，验证集 $S_{validate}$，群体规模 n，交叉概率 p_c，变异概率 p_m，最大进化代数 T

输出：最优特征子集 F_{select}

1. 初始化，随机生成群体 P，$|P|=n$；
2. 运行 NSGA-II 算法 nsga2r(S_{train}, S_{test}, n, T, p_c, p_m)，直至收敛；
3. 生成 Pareto 最优解集 $P_{optimal}$；
4. $F_{select}\leftarrow\Phi$，重复以下步骤，直至 $P_{optimal}$ 为空：
   4a. 运行 $SVM(F,\ S_{validate})$，$F\in P_{optimal}$，若 $SVM(F_{select},\ S_{validate})\pi SVM(F,S_{validate})$，则 $F_{select}=F$；
   4b. $P_{optimal}\leftarrow P_{optimal}-\{F\}$；
5. F_{select} 即为最优特征子集；

图 1 基于多目标遗传算法的特征选择算法

5 实验与分析

5.1 数据集描述

为了准确地对特征选择算法进行评价，本文选择了 KDD CUP 1999 数据集[10]作为测试集。KDD CUP 1999 数据集是在入侵检测领域广泛使用的测试集，采用 1998 DARPA 入侵检测数据集来构造连接记录及提取特征。数据中包含 41 个特征，主要分为 4 种攻击类型：(1)DoS，拒绝服务攻击；(2)R2L，对远程主机的未授权的访问；(3)U2R，对本地超级用户权限的未授权的访问；(4)Probe，扫描与探测行为。

由于 KDD 原始数据集过于庞大，为了对算法进行验证，本文选择了其中具有代表性的 corrected 数据集，分别随机生成训练集、测试集与验证集，其中分别包含了 5072、5280 与 10158 条记录。在测试中，取 10 次运行算法的平均值作为测试结果。

5.2 基于 Pareto 的多目标遗传算法的实验与分析

基于 Pareto 的方法能够克服传统的多目标优化方法中权值矢量的影响。另外，由于 NSGA-II 算法采用了小生镜技术来保持群体的多样性，因此避免了“过早收敛”到特定区域的问题。这些方法使得对解空间的搜索趋向 Pareto 最优前沿。图 2 显示了群体在演进过程中的变化趋势。

从图 2 可以看出，在特征数量与检测精度构成的目标空间中，虽然数据点比较分散，但最终趋向于 Pareto 最优前沿。同时，我们注意到，当特征数目分布在[13, 27]区间时，部分解的 1-Fvalue 值聚集在 0.06 至 0.07 之间，这说明在入侵检测模型中，特征数量与检测精度之间并没有必然的联系，较多的特征并不能产生更好的检测精度，从而表明了需要在特征数量与检测精度之间寻求最优是必要的。

另外，为了在 Pareto 最优解集中选择最优的特征子集，我们利用了单独的验证集，通过交叉验证来对比测试集与验证集的生成结果，选择泛化精度最高的特征子集。Pareto 最

维普资讯 http://www.cqvip.com

优前沿与验证集获得的曲线对比如图3所示。

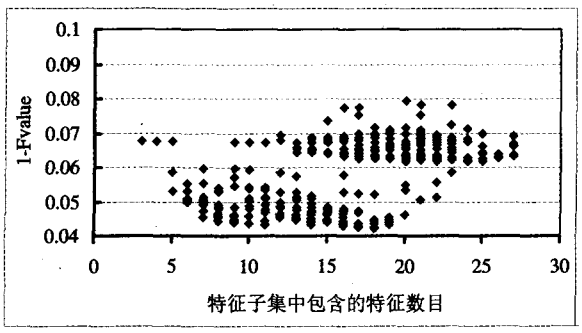

图2 基于Pareto方法的特征子集群体演进趋势

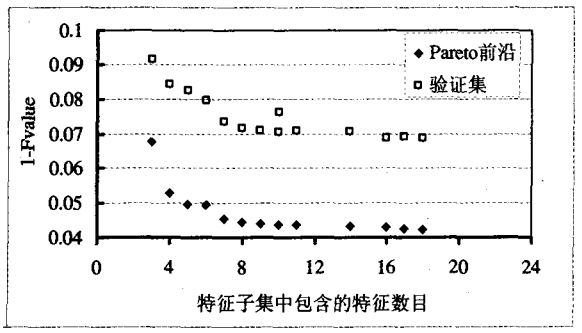

图3 Pareto最优前沿与验证集曲线对比

从图3中，可以看出，当特征子集的基数为18时，无论是Pareto最优前沿还是由验证集生成的曲线，均具有较好的检测精度，因此我们选择特征数目为18的特征子集作为最终的最优特征子集。同时，我们还注意到，当特征数目为17时，Pareto最优前沿与验证集曲线中检测精度的表现是不同的，即在Pareto前沿中，其精度优于特征数目为16时的情况，而在验证集曲线中则相反。由此可见，采用验证集在Pareto最优前沿中寻求泛化精度最高的特征子集是必要的。表1列出了此时最优特征子集中包含的特征。

5.3 入侵检测精度的实验与分析

为了验证根据多目标优化遗传算法所得到的特征子集构建的入侵检测模型的性能，我们分别利用所得到的特征子集与原始特征集构建入侵检测模型，测试其检测结果。二者检测结果的对比如表2所示。

表2 原始特征集与优化特征集的检测率与误报率比较

	原始特征集	优化特征集
特征数量	41	18
检测率(DR%)	98.574%	99.216%
误报率(FPR%)	9.167%	7.479%

由表2可以看出，优化后的入侵检测模型，其检测率与误报率均略好于原始的入侵检测模型，这说明利用优化特征子集构建的入侵检测模型不但显著提高了检测速度，而且其精度也略有上升。因此，基于多目标遗传算法的特征选择方法是有效的，能够提高入侵检测的性能。

结论 针对入侵检测中存在着不相关或冗余的特征从而导致检测性能下降的问题，本文提出了一种面向入侵检测的基于多目标遗传算法的特征子集选择方法，只选择最具分类信息的特征构建入侵检测模型。在该方法中，根据入侵检测的特点，将特征选择问题作为多目标优化问题来考虑，充分利用遗传算法的全局搜索及并行处理能力，并采用基于Pareto优化的方法，生成Pareto最优解集。同时，为了避免"过拟和"问题，采用了交叉验证方法，从Pareto最优前沿中选择泛化精度最好的解作为最终的入侵检测特征集合。试验结果显示，在显著提高算法效率的基础上，其检测精度也高于采用原始特征集合的入侵检测模型，有效地提高了入侵检测系统的性能。

致谢 感谢印度Kanpur遗传算法实验室的Kalyanmoy Deb博士提供了NGSA-II算法的源代码。

表1 最优特征子集

序号	数据特征	序号	数据特征	序号	数据特征
1	duration	12	logged_in	34	dst_host_same_srv_rate
2	protocol-type	13	num_compromised	35	dst_host_diff_srv_rate
3	service	17	num_file_creations	36	dst_host_same_src_port_rate
4	flag	22	is_guest_login	39	dst_host_srv_serror_rate
5	src_bytes	31	srv_diff_host_rate	40	dst_host_rerror_rate
6	dst_bytes	32	dst_host_count	41	dst_host_srv_rerror_rate

参考文献

1 Liu Huan, Yu Lei. Toward Integrating Feature Selection Algorithms for Classification and Clustering. IEEE transactions on Knowledge and Data Engineering, 2005,17(3):491～502

2 Lee Wenke, Stolfo S J, Mok K W. A Data Mining Framework for Building Intrusion Detection Models. In: Proceedings of the 1999 IEEE Symposium on Security and Privacy, Oakland, California, May 1999

3 Sung A H, Mukkamala S. Identifying Important Features for Intrusion Detection Using Support Vector Machines and Neural Networks. In: Proceedings of International Symposium on Applications and the Internet (SAINT 2003), 2003. 209～217

4 Chebrolu S, Abraham A, Thomas J P. Hybrid Feature Selection for Modeling Intrusion Detection Systems. In: the 11th International Conference on Neural Information Processing (ICONIP04), 2004

5 Fonseca C M, Fleming P J. An Overview of Evolutionary Algorithms in Multiobjective Optimization. Evolutionary Computation, 1995,3(1):1～16

6 Goldberg D E. Genetic Algorithms in Search, Optimization and Machine Learning. New York: Addison Wesley, 1989

7 Deb K, Pratap A, Agarwal S, et al. A Fast and Elitist Multi-objective Genetic Algorithm: NSGA-II. IEEE Transcation on Evolutionary Computation, 2002, 6(2):181～197

8 Mukkamala S, Sung A H, Abraham A. Intrusion detection using an ensemble of intelligent paradigms. Journal of Network and Computer Applications, 2005, 28(2): 167～182

9 Joshi M, Agarwal R, Kumar V. Predicting Rare Classes: Can Boosting Make Any Weak Learner Strong?. In: Proceedings of Eight ACM Conference ACM SiGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, Canada, 2002

10 The UCI KDD Archive. KDD99 cup dataset. http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html, 1999