文章编号：1006－9348(2008)09－0115－03

基于遗传算法的 SVM带权特征和模型参数优化

杨 洁[1]，郑 宁[1]，刘 董[1]，罗时贵[2]

（1. 杭州电子科技大学计算机学院，浙江 杭州 310018；2 中国人民银行台州市中心支行，浙江 台州 318000）

摘要：建立在统计学习理论和结构风险最小原则上的支持向量机（SVM）在理论上保证了模型的最大泛化能力，因此将支持向量机理论应用于入侵检测领域可以获得很好的效果。但是在应用中也存在如何对网络数据进行特征编码和选择适当的支持向量机模型参数的问题。在分析了特征编码和模型参数对分类器识别精度的影响基础上，提出用遗传算法建立支持向量机带权特征和分类器模型参数的自适应优化算法，并在网络入侵检测中成功的运用算法。最后，使用 KDD CUP 1999数据进行的仿真实验表明了算法的正确有效性。

关键词：遗传算法；支持向量机；入侵检测

中图分类号：TP309 文献标识码：A

Optimization of Features with Weight and Model Parameters of SVM Based on Genetic Algorithm

YANG Jie, ZHENG Ning, LIU Dong, LUO Shi-gui

(1. College of Computer, Hangzhou Dianzi University, Hangzhou Zhejiang 310018, China;
2. Taizhou Center Sub-branch, The People's Bank of China, Taizhou Zhejiang 318000, China)

ABSTRACT: The support vector machine (SVM), which is based on the statistical learning theory and the structural risk minimum principle, guarantees the largest generalization ability of a model. So it can obtain very excellent effect when using the SVM theory in the field of intrusion detection. However, in the application it also has some problems such as how to code the features of network data and how to select the proper model parameters of SVM. The paper proposed a self-adaptive optimization algorithm for the SVM feature with weight and model parameters using genetic algorithm on the basis of analyzing the influence of feature coding and model's parameters on the classifier's recognition accuracy. And the algorithm was successfully applied for network intrusion detection. Finally, experiment shows the effectiveness of the optimization algorithm using the KDD CUP 1999 Data.

KEYWORDS: Genetic algorithm; Support vector machine (SVM); Intrusion detection

1 引言

支持向量机是由 Vapnik等人于 1995年提出的从统计学习理论发展出的一种模式识别方法[1]。它建立在统计学习理论和结构风险最小的基础上，在理论上充分保证了模型的泛化能力，目前已经广泛运用于模式识别领域[2]。入侵检测可看成是一个对网络数据进行分类的问题。网络数据十分复杂，常常体现为高维、线性不可分等特点。与传统的学习方法相比，支持向量机具有小样本、良好的推广性能、全局最优等优点[3]，所以支持向量机方法可以很好地解决这类问题。将支持向量机理论应用于入侵检测领域目前已经成为一种趋势。

在基于 SVM的网络入侵检测系统中，支持向量机训练模型有许多参数要进行选择，比如核函数的选取及核函数的相关参数等。这些参数直接影响 SVM的分类能力。同时，特征属性的编码方式也对 SVM的分类能力和应用范围有很大的影响。如何选择合适的训练模型参数和特征属性编码方式，目前还没有统一的标准。实际上，特征属性的编码和训练模型参数的优化过程是互相依赖的，因为优化其中一个时需要确定另一个才能进行测试。然而在目前的应用中，它们却常分开进行或者只对其中的一个进行优化。这样做存在无法确定先优化特征属性编码还是训练模型参数的缺陷。因此为了使 SVM有更好的性能，特征属性的编码和训练模型参数的优化应该同时进行。本文通过研究分析训练模型参数和特征属性编码对 SVM的性能影响，提出了基于

基金项目：浙江省自然科学基金（Y106176）；浙江省科技厅科技计划项目（2007C33058）

收稿日期：2007－08－14 修回日期：2007－10－09

遗传算法的模型参数和带权特征的自适应共同优化方法。该方法在训练过程中将特征属性的权值和模型参数的获取通过遗传算法进行共同优化，获得适当的模型参数和带权特征编码方式。

2 支持向量机

2.1 线性情况

假设一组训练样本 $D=\{(x_i, y_i)\}_{i=1}^{n}$，其中，$x_i\in R$，$y_i\in\{-1, 1\}$，$R$表示输入模式的特征空间。当训练样本集线性可分时，可以将分类超平面的描述为 $w\cdot x+b=0$，其中向量 w 为分类超平面的权系数。b是分类阈值。要找到这个超平面，需要求解下面的二次规划问题：

$$\min\phi(w)=(w\cdot w)/2$$
$$s.t\quad y_i[(x_i\cdot w)-b]\geq 1,\quad i=1, 2, \cdots, n \qquad (1)$$

此时，可以求解得到最优分类超平面的线性分类器为

$$f(x)=\operatorname{sgn}\left[\sum_{i=1}^{n} a_i y_i (x_i\cdot x)+b\right] \qquad (2)$$

其中 a_i为拉格朗日乘子，其中不为 0的解向量为支持向量。

2.2 非线性情况

对于非线性问题，支持向量机通过选择适当的非线性变换，将输入空间 R中的训练样本映射到某个高维特征空间 F，使得在目标高维空间中这些样本线性可分。

根据泛函的有关理论，若核函数 $K(x, x_i)$ 满足 Mercer条件，它就对应某一变换空间中的内积 $<\phi(x_i\cdot x)>$，函数 $\phi: R\rightarrow F$是一个从非线性输入空间 R到高维特征空间 F的映射，所以求映射 $\phi: R\rightarrow F$只要知道如何由输入 x, x_i计算内积 $<\phi(x_i\cdot x)>$即可，由

$$K(x, x_i)=\phi(x_i)\cdot\phi(x) \qquad (3)$$

将式(2)重写，即可得到对应高维空间的非线性分类器为

$$f(x)=\operatorname{sgn}\left[\sum_{i=1}^{n} a_i y_i K(x_i\cdot x)+b\right] \qquad (4)$$

这样，在高维空间的内积运算转化为低维输入空间中的一个简单函数计算。

采用不同的核函数将导致不同的支持向量计算法，目前广泛应用的核函数形式主要有线性核函数、多项式核函数、径向基核函数(RBF)、Sigmoid核函数。由于径向基函数在基于 SVM的入侵检测中应用比较多，因此，本文选取径向基函数作为和支持向量机的核函数来进行研究。径向基函数为[4]

$$K(x, x_i)=\exp\{-\eta\|x-x_i\|^2\} \qquad (5)$$

3 特征编码和模型参数对分类能力的影响

3.1 SVM的分类能力评价

SVM的分类能力是指其对新数据的分类识别能力。对于两分类问题，分类能力通常使用分类器对测试数据的识别精度(Recognition Accuracy, RA)来表示(又称识别率，检测率)，定义如下：

$$RA=\frac{\text{测试数据集中分类正确的个数}}{\text{测试数据集样本个数}} \qquad (6)$$

3.2 特征编码对 SVM性能的影响和优化

SVM训练和检测的对象都是标准数据样本，但是特征编码过程会直接影响标准数据样本的形成，可见特征编码过程会影响到 SVM分类器的识别精度。数据采集过程中往往采集的特征数都比较多，而且各个特征的取值范围相差有可能很大，如果不进行处理，取值范围大的特征属性掩盖了取值范围小的特征属性的作用。目前对特征属性进行处理主要是特征的选择和特征属性归一化，即选择部分相关特征，并把这部分数据处理到一个相同的范围内，如[-1, 1]。但是，在实际 SVM的运用中，各个特征属性的作用不一定等价，它们对分类器识别精度的影响也不尽相同。因此，统一归一化处理的过程虽然看上去公平的对待了所有的特征属性，实际上降低了那些对分类能力影响大的特征属性的作用。同时，目前对特征的选择并无定论，大多是凭经验数据。在考虑到这方面的问题后，本文在归一化的过程中引入模糊的思想，提出了一种带权的特征属性标准化方法。

对于特征集 $F=\{f_1, f_2, f_3, \cdots, f_n\}$，其中 n表示特征集中的特征数目，如果对其中的每一个特征 $f_i(i=1, 2, \cdots, n)$用 w_i $w_i\in[0, w_{max}]$ 来表示它的权值，w_{max}是正常数，那么带权特征编码可以用一个 n位的实数序列表示为：$W=\{w_1, w_2, w_3, \cdots, w_n\}$，$w_i\in[0, w_{max}]$，$i=1, 2, 3, \cdots, n$其中每一位 w_i用来表示对应的 f_i的权值。用 SVM的分类器识别精度 RA作为目标函数，带权特征的优化问题可以描述如下：

$$\max RA(W) \qquad (7)$$
$$s.t\quad W=\{w_1, w_2, w_3, \cdots, w_n\},\ w_i\in[0, w_{max}],\ i=1, 2, 3, \cdots, n$$

如果用 lower和 upper分别表示预先设定的所有特征值进行归一化处理的下限和上限，那么对于每个特征 $f_i(i=1, 2, 3, \cdots, n)$，它标准化后的取值范围是 $[w_i\cdot lower, w_i\cdot upper]$，权值的差异也就体现出了特征属性之间对分类器识别精度影响上的差异。当 w_i全都取 0或者 w_{max}的时候，优化后得到权值序列实际上也就是个特征选择序列(0表示不选择，w_{max}表示选择)，可见特征选择是带权特征的一个特例。从上面的分析可以知道，带权特征实际上表示的是一种特征编码的方法，既包括特征选择，也包括了数据的标准化处理。如果用演化算法求解该问题，那么可以发现在计算 RA前要有确定的 SVM的训练模型及其模型参数。

3.3 模型参数对分类精度的影响及优化

在实际的使用中，对分类精度有重要影响的几个参数主要是：惩罚因子 C，核函数的参数。SVM用惩罚因子 C调整最小经验风险和置信度的折中，从而达到最小实际风险。C越大则对数据的拟合程度越高，但泛化能力将降低[6]。核函数的相关参数，比如多项式核函数的多项式次数；RBF的 η值对模型的分类精度均有重要影响。由于 RBF径向基函数对非

线性、高维数据有很好的适应性，在基于 SVM的网络入侵检测系统中的经常被采用，所以本文也选定 RBF作为核函数，主要对其 值进行优化研究。结合 RBF的参数 和惩罚因子 C，可以得到需要优化的组合参数 P= { C, }，其中 C> 0, > 0。为了寻找支持向量机分类器的最佳模型参数。用 SVM的分类器识别精度 RA作为目标函数，模型参数的优化问题可以描述如下：

$$\max \mathrm{RA}(P) \qquad (8)$$
$$\text{s.t} \quad P = \{C, \gamma\}$$
$$C > 0, \ \gamma > 0$$

如果用演化算法求解该问题，不难发现，求解过程中要预先确定特征编码后才能获得进行 SVM训练和分类操作所需要的标准数据，从而计算出 RA。目前，在应用中对模型参数优化的时候，大多是凭经验选择其中的一些特征[5]，也有一些选择了全部的特征[6]（此时特征数较少）来进行编码。

3.4 带权特征和模型参数的共同优化方法

通过上面的分析可知，对于使用 RBF为核函数的 SVM，带权特征和模型参数的优化具有相互依赖性。即在优化带权特征 W时，要预先确定一个模型参数 P；在优化模型组合参数 P前，要预先确定一个带权特征 W。实际上式（7）和（8）的优化目标函数都是提高 SVM分类器识别精度 RA。因此本文提出特征选择和模型参数的共同优化，可以描述如下：

$$\max \mathrm{RA}(W, P) \qquad (9)$$
$$\text{s.t}$$
$$W = \{w_1, w_2, w_3, \cdots, w_n\}, \ w_i \in [0, w_{\max}], \ i = 1, 2, 3, \cdots, n$$
$$P = \{C, \gamma\}, \ C > 0, \ \gamma > 0$$

其中混合参数中的前一个参数为特征选择，后一个参数为 RBF核 SVM的模型参数。式（9）同时对特征选择和 SVM模型参数进行优化，从而解决了式（7）和（8）中由于特征选择和模型参数的相互依赖导致的先后次序问题。由于遗传算法（GA）[6] 具有隐含的并行性和强大全局搜索能力，可以在很短的时间内搜索到全局最优点。因此本文采用 GA来求解。

为了使得 SVM具有更好的性能，进行实验的数据要进行标准化处理。如果将所有特征都归一化处理到[lower, upper]范围内，那么对于每个特征 f_i（$i = 1, 2, \cdots, n$），需要将其数据的范围约束到[w_i。lower, w_i。upper]。对于连续型特征属性，标准化处理方法如下：

$$d_i = w\left(\text{lower} + \frac{(\text{upper} - \text{lower})(y_i - \min f_i)}{(\max f_i - \min f_i)}\right)$$
$$i = 1, 2, 3, \cdots, n \qquad (10)$$

其中 d_i表示第 i个特征属性归一化后的标准数值，y_i表示第 个特征属性的原始数值，$\min f_i$表示第 i个特征属性的最小值，$\max f_i$表示第 i个特征属性的最大值。对于字符型特征属性，按照最大间距的原则转换成其取值范围内的不同的离散数值。

对于选定的样本数据集，要先随机同分布的选出 10%的数据作为原始训练数据，随后再把原始训练数据分成两部分 Data_1、Data_2。Data_1用于 SVM的训练，Data_2用于验证，每部分各占 50%。

使用 GA进行处理时，个体为 SVM的带权特征 W和模型组合参数 P，其适应度为 SVM分类器的识别精度 RA。首先，对每个个体进行二进制编码，随机产生初始化种群；其次，解码种群中的染色体，获取 W和 P的值，根据 W的值对样本数据进行编码；再次，使用参数 P运用 Data_1训练 SVM分类器模型，并用训练好的 SVM分类器计算 Data_2的识别精度 RA，然后判断遗传算法的停止准则是否满足，如果满足则停止计算，输出最优参数，否则，则执行选择、交叉和变异等操作以产生新一代种群，并开始新一代的遗传。该方法是在 SVM进行训练前自适应的进行带权特征和模型参数优化，一旦得到优化结果后就不需要再次进行优化，因此该方法会占用较多的编码训练时间，但是不会增加 SVM检测的时间。

4 仿真实验

4.1 数据源

实验中采用的训练数据和测试数据来自 DARPA提供的一个网络流量异常检测的标准数据集（KDD Cup 1999 Data），它是由美国麻省理工学院的 Lincoln实验室建立的，从实际网络获得原始的 TCP/IP网络通信数据，对每一个 TCP/IP连接记录提取出了 41个特征[5]。每个特征可能的取值又不相同，是典型的异构数据集。该数据集包括约 500万条连接记录和 300万条测试记录，其中有大量的正常网络流量和各种攻击，具有很强的代表性，训练数据带有标记（正常或某种攻击）。数据中包括四种类型的攻击和正常数据：DoS（拒绝服务攻击）、R2L（远程权限获取）、U2R（对本地超级用户的非法访问）、Probe（各种端口扫描和漏洞扫描）和 Normal（正常数据）。

由于数据集过于庞大，本文从中使用了两组数据分别进行实验，一组是训练集中 Probe和 Normal数据组成的数据集（简称 Probe数据集），另一组是 10%训练数据集（简称 10%数据集）。异常和正常的数据被分别标志为“1”和“-1”。SVM使用 LIBSVM[5]的代码。

4.2 实验结果及分析

在本文的遗传算法中，交叉率和变异率分别为 0.7和 0.04，数据归一化范围[lower, upper]取为[-1, +1]，$w_{\max} = 7/8$。染色体采用二进制编码，总长度为 133，其构成方式如图 1所示。W由 123位二进制编码构成，每三位构成一个 w_i，$i = 1, 2, 3, \cdots, 41$。P由 10位二进制编码构成，前 2位构成 C，后 8位构成 。解码方法为：$w_i = w_i/2^3$，$i = 1, 2, 3, \cdots, 41$

$$C = 10^{C-1}, \quad r = \left(\frac{r+1}{100}\right)^2$$

使用 GA对特征选择和模型参数进行共同优化，遗传进化 20代后，交叉验证的结果如图 2所示。经过 20代遗传后，SVM对两种数据集都达到了很高的识别精度，取得了很好的优化结果。Probe类型数据集中的 （下转第 158页）

参考文献:

[1] G J Foschini and M J Gans. On limits of wireless communications in a fading environment when using multiple antennas[J]. Wireless Personal communications, 1998, (6): 311 - 335.

[2] A Grant. Rayleigh fading multiple-antenna channels[J]. EURASIP Journal on Applied Signal Processing, March 2002, 2002 (3): 316-329.

[3] David Tse Pramod Viswanath. Foundamentals of Wireless Communication[M]. 北京: 人民邮电出版社, 2007.

[4] 陈其铭. 无线 MIMO系统的传输性研究[D]. 广东: 华南理工大学博士学位论文, 2005.

[5] 金荣洪, 耿军平, 范瑜. 无线通信中的智能天线[M]. 北京: 北京邮电大学出版社, 2006.

[6] 高惠璇. 应用多元统计分析[M]. 北京: 北京大学出版社, 2003

[作者简介]

袁 峰 (1977—), 男 (汉族), 陕西人, 硕士研究生, 主要研究方向: 移动通信, 通信信号处理;

张 捷 (1964—), 男 (汉族), 安徽滁州人, 西北工业大学电子信息学院副教授, 主要研究方向: 移动通信, 通信信号处理等。

(上接第 117页)

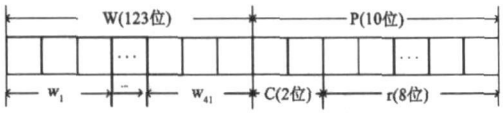

图 1 染色体的构成方式

攻击种类比 10%训练数据集少很多, 因此 probe类型数据在优化过程中, 识别率明显高于 10%数据集的原因。

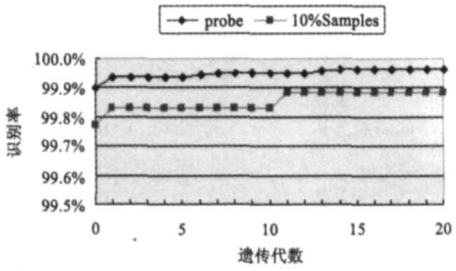

图 2 使用 GA优化的效果

最终优化得到的 W, P值和对测试样本的检测率见如表 1所示。优化 20代后, 虽然 Probe数据集包含的攻击种类比 10%数据集少, 但是 Probe数据集的特征权值较大(不小于 0.5)的数目比 10%数据集的特征权值较大的数目还要多 4个。这说明数据样本的复杂度增加并不能使得特征权值序列里较大特征值的数目一定增加。这与特征选择中复杂样本选择的特征属性往往比简单样本选择的多不一样。

表 1 使用 GA优化得到的参数和结果

	$w_i \geq 0.5$	P(C, r)	RA
Probe数据集	21个	100, 0.485	99.9601%
10%数据集	17个	10, 0.325	99.8822%

实验结果显示, 在网络入侵检测中使用 GA对对 SVM带权特征和模型参数进行共同优化可以取得很好的识别率。在攻击类型单一的情况下, 该方法的 RA非常好, 因此在对网络数据进行分类检测时, 该方法的优势将更加明显。

5 结论

本文提出了一种基于遗传算法的支持向量机带权特征和模型参数优化方法, 并把此方法应用到网络入侵检测中。它避免了 SVM传统模型参数优化需要依赖于经验数据进行特征编码的缺陷, 具有更好的自适应性和数据编码方式, 以及样本识别精度高的良好表现。进一步的工作主要集中在: 对网络攻击特征的抽取、核函数的研究等方面, 进一步缩短检测时间, 提高检测精度。

参考文献:

[1] Vapnik V. The Nature of Statistical Learning[M]. New York: Springer, 1995

[2] Ge M, Du R, Zhang G C, Xu Y S. Fault diagnosis using support vector machine with an application in metal stamping operations [J]. Mechanical Systems and Signal Processing, 2004, 18: 143 ~159.

[3] 沈翠华, 刘广利, 邓乃扬. 一种改进的支持向量分类方法及其应用[J]. 计算机工程, 2005, 31(8): 153 - 154.

[4] 李 青, 焦李成, 周伟达. 基于向量投影的支撑向量预选取[J]. 计算机学报, 2005, 28(2): 145-152

[5] 付长龙, 吕彦波, 姚全珠, 杜旭辉. 基于样本密度的 SVM及其在入侵检测中的应用[J]. 计算机应用. 2007, 27(4): 838-840.

[6] 陈 果. 基于遗传算法的支持向量机分类器模型参数优化[J]. 机械科学与技术, 2007, 26(3): 347-350

[作者简介]

杨 洁 (1984—), 男 (汉族), 湖南岳阳人, 硕士研究生, 研究方向: 网络安全, 入侵检测;

郑 宁 (1961—), 男 (汉族), 浙江杭州人, 博士生导师, 教授, 研究方向: 计算机应用;

刘 董 (1982—), 女 (汉族), 浙江舟山人, 硕士研究生, 研究方向: 网络安全, 入侵检测;

罗时贵 (1982—), 男 (汉族), 浙江台州人, 助理工程师, 研究方向: 入侵检测。