基于支持向量机的入侵检测系统的优化

optimization in building an intrusion detection system based on support vector machine

简清明 （四川理工学院网管中心 643000）

摘要：本文关注基于支持向量机算法的入侵检测系统的优化问题。首先，介绍一个简单的基于单 SVM 的入侵检测系统。然后，推荐一种对输入特征进行重要性排序的方法。最后，根据分类结果，提出基于多 SVMs 的入侵检测系统模型。

关键词：入侵检测系统 网络安全 支持向量机 模式识别 重要特征

和传统的入侵检测系统相比，基于支持向量机的入侵检测系统具有较少的训练时间，并且在先验知识不足的情况下，仍有较好的分类正确率。下面我们介绍一种常用的对特征信息重要性进行排序的方法和随后的基于排序结果的基于支持向量机的入侵检测系统的优化模型及其性能变化。

1 实验说明

我们这里所使用的实验数据全部源于美国麻省理工学院的林肯实验室。这是由 DARPA（美国国防部高级研究计划局）为评估入侵检测而发展起来的，被认为是评估入侵检测系统性能的基准。现在有两个公开的数据集提供下载——1998 和 1999。这里采用 1998 年的评估结果。在 1998 DARPA 入侵检测评估中[2]，实验室研究人员通过模拟一个典型的美国空军网络以获得原始的 TCP/IP 网络通信数据。对每一个 TCP/IP 连接，41 个不同数量和性质的特征被提取．我们从中选取一个数据子集，包含 494,021 的数据记录，其中 20% 为正常模式。

所有的攻击分为四个主要类别：

(1) Probe：监视和其他的窥探行为；

(2) DOS：拒绝服务攻击 ；

(3) U2Su：对本地特权用户的非授权访问；

(4) R2L：源于远程主机的对本地服务器的非授权访问。

后面描述的所有基于 SVMs 的实验都采用免费的软件包 SVM light 完成。可以在 http://www.cs.cornell.edu/People/tj/svm_light/下载它的最新版本。

2 对输入的重要性进行排序的方法

对特征信息的分类选择面临和其他的工程项目相同的问题，表现在以下三个方面：

(1) 有不同重要程度的大量输入变量 x = (x1, x2, …, xn)，其中一些是本质因素，一些是一般因素，另外则是不相关的无用的噪音。

(2) 缺乏分析模型或数学公式以精确描述输入输出关系 y = F (x)。

(3) 可在有限的数据集上构建模型用于模拟或预测。

因此，我们采用一次去除一个特征并使用 SVMs 检测随之的性能变化的方法来对输入特征进行排序。步骤如下[1]：

① 编排训练集和测试集；对每一个特征进行以下操作

② 从训练集和测试集中删除此特征；

③ 使用处理后的数据集训练分类器；

④ 根据选定的性能标准，使用测试集分析分类器的性能表现；

⑤ 根据规则对特征的重要性进行分类。

对基于支持向量机的入侵检测系统的输入特征信息的重要性进行排序，我们考虑三个主要性能指标：分类精确度（5 类，包括一类正常模式和四类攻击模式）、训练时间和测试时间。用于分类的规则和规则集如下[1]：

* 如果精确度降低而训练时间或测试时间之一增加，或精确度不变而训练时间和测试时间都增加，则该特征为重要特征。

* 如精确度不变而训练时间和测试时间之一增加，或精确度增加而训练时间和测试时间之一增加，另一减少，则该特征非重要特征或一般特征。

* 如精确度不变而训练时间和测试时间都减少，或精确度增加而训练时间和测试时间都减少，则该特征无关紧要特征。

根据上述规则，41 个特征分成三类{重要}，<一般>，和（无关紧要），对五类模式的测试结果如下：

** 正常(Normal)：{1,3,5,6,8－10,14,15,17,20－23,25－29,33,35,36,38,39,41}，<2,4,7,11,12,16,18,19,24,30,31,34,37,40>，(13,32)

** Probe：{3,5,6,23,24,32,33}，<1,4,7－9,12－19,21,22,25－28,34－41>，(2,10,11,20,29,30,31,36,37)

** DOS：{1,3,5,6,8,19,23－28,32,33,35,36,38－41}，<2,7,9－11,14,17,20,22,29,30,34,37>，(4,12,13,15,16,18,19,21,3)

** U2Su：{5,6,15,16,18,32,33}，<7,8,11,13,17,

19－24,26,30,36－39>,(9,10,12,14,27,29,31,34,35,40,41)

＊＊ R2L:{3,5,6,24,32,33},<2,4,7－23,26－31,34－41>,(1,20,25,38)

分析上面的结果,我们可以发现:正常模式有25个重要特征,Probe只有7个,DoS有19个,U2SU有8个,R2L最少只有6个,相对于全部41个特征而言,都有大幅度的减少,特别是Probe、U2SU、R2L.这意味着一旦确定了输入特征的重要程度,我们就可以只采用重要特征来训练和测试分类器,这将大大提高入侵检测系统的处理性能和检测效率。随后我们通过实验比较了使用全部特征、只使用重要特征以及重要特征和一般特征的集合的性能表现。实验的结果证实了我们的推断。

3 新的基于SVMS的入侵检测系统模型

根据分类的结果,我们首先构建一个新的基于多SVMS的入侵检测系统模型(如图1所示)。

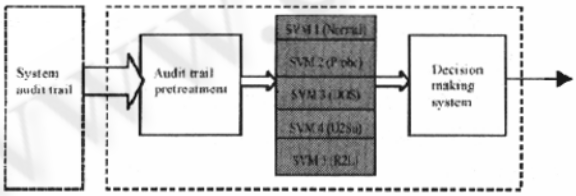

图 1　采用五个SVMS的入侵检测系统模型

该模型使用了五个分类器。这是因为,SVMs只能进行二元分类,而我们的网络通信模式分为五个行为集合,所以将使用五个SVMs来构建IDS,每个SVM负责一类,对应五个行为集合.由于五个分类中的重要特征是不完全相同的,从而使用五个SVMs成为一个优点而不是缺点——它可以有效地减少所需处理的输入信息,从而提高IDS的性能。

现在来看一看这个模型工作的情况。这里仍然使用SVM－light来模拟每个SVMS的分类操作。最终的实验结果如下:

表 1　使用全部41个特征时svms表现

类别	训练时间(S)	测试时间(S)	精度(%)
Normal	7.71	1.27	99.54
Probe	48.97	2.09	99.72
Dos	22.85	1.93	99.27
U2su	3.42	1.06	99.89
R2l	11.65	1.01	99.79

表 2　使用重要特征SVMS的表现

类别	特征数目	训练时间(S)	测试时间(S)	精度(%)
Normal	25	9.41	1.08	99.58
Probe	7	39.65	1.87	99.40
Dos	19	22.81	1.86	99.23
U2su	8	2.62	0.89	99.87
R2l	6	8.79	0.81	99.78

表 3　使用重要和一般特征时SVMS的表现

类别	特征数目	训练时间(S)	测试时间(S)	精度(%)
Normal	39	8.25	1.21	99.59
Probe	32	46.98	2.07	99.65
Dos	32	20.67	2.03	99.25
U2su	25	2.81	0.94	99.87
R2l	37	8.36	1.22	99.80

4 结论

观察上面三个表,可以发现:SVMs很容易达到很高的精度(超过99%),与使用的特征是全部特征、重要特征还是重要特征与一般特征的集合没有关系——其精度差异在统计上可以忽略不计。换句话说,在去掉一般输入特征信息和无用的信息后,基于SVMS的IDS的探测精度不受丝毫影响。相反,只使用重要特征还带来非凡的性能提升:测试时间在每个类都减少,Normal的检测精度有轻微上升,Probe和DoS有轻微减少,U2SU和R2l则保持不变。

参考文献

1 Srinivas Mukkamala. Audit Data Reduction Using Neural Networks and Support Vector Machines [EB/OL]. http://www.dfrws.org/agenda－msw－1－3.htm.2002－08－08 .

2 1998 DARPA Intrusion Detection Evaluation [EB/OL]. http://www. ll. mit. edu/IST/ideval/docs/docs_index.html.

3 饶鲜,基于支持向量机的入侵检测系统[EB/OL]. http://www.jos.org.cn/1000－9825/14/798.htm. 2001－12－10.

4 Srinivas Mukkamala & Andrew H. Sung. Identifying Significant Features for Network Forensic Analysis Using Artificial Intelligent Techniques [EB/OL]. http://citeseer. nj. nec. com/585832. html. 2003.