基于网格搜索的 SVM 在入侵检测中的应用

张公让,万 飞
(合肥工业大学 管理学院,安徽 合肥 230009)

摘 要:随着网络的快速普及和发展,网络安全问题日益突出,如何保障网络安全已经成为一个国际化问题。在众多方法中,入侵检测技术是解决这一问题的有效手段。文中将支持向量机方法运用在入侵检测中。首先,介绍了基于 SVM 的入侵检测技术研究现状;然后,将网格搜索算法应用在 SVM 参数寻优中;最后,通过实验,将 PSO 算法、GA 算法、网格搜索算法对 SVM 参数优化的结果进行比较。实验结果表明,使用网格搜索法对 SVM 参数进行优化,具有最好的泛化精度,并且在此基础上,对数据集进行归一化处理,将大幅度减少构建分类器的迭代次数,从而减少预测时间。因此,可以认为基于网格搜索的支持向量机能够很好地实现入侵检测。

关键词:入侵检测;网络安全;支持向量机;网格搜索
中图分类号:TP39 **文献标识码**:A **文章编号**:1673-629X(2016)01-0097-04
doi:10.3969/j.issn.1673-629X.2016.01.020

Application of Support Vector Machine in Network Intrusion Detection Based on Grid Search

ZHANG Gong-rang, WAN Fei
(School of Management, Hefei University of Technology, Hefei 230009, China)

Abstract: With the rapid popularization and development of network, network security problems are becoming increasingly prominent. How to guarantee the security of the network has become an international problem. Among the many methods, intrusion detection technology is an effective means to solve this problem. In this paper, Support Vector Machine (SVM) method will be used in intrusion detection. First of all, the current situation of the intrusion detection technology is introduced based on SVM. Secondly, the grid search algorithm is used into the optimization of the SVM's parameters. At last, bring the result of the SVM's parameters that based on PSO algorithm, GA algorithm and grid search algorithm into comparison. The results of the experiment show that using the grid search method for optimization of SVM's parameters has the best generalized accuracy, and on this basis, the normalization of dataset will greatly reduce the number of the classifier's iterations, so as to reduce the forecast time. Therefore, it is considered that SVM based on grid search can realize the intrusion detection excellently.

Key words: intrusion detection; network security; support vector machine; grid search

1 概 述

网络入侵检测是指从计算机网络的若干关键点收集信息并对其进行分析,从中查找网络中是否有违反安全策略的行为或遭到入侵的迹象,并依据既定的策略采取一定的软件与硬件的组合措施予以防治[1]。1980 年,James Anderson 为美国空军做的技术报告《Computer Security Threat Monitoring and Surveillance》里面第一次提出入侵检测的理念,他将入侵检测类型划分为外部入侵、内部用户的越权限使用和授权用户的权限滥用三种,并提出用审计踪迹来检测对文件的非授权访问。1987 年,Dorothy. E. Denning 首次实现了一个实时入侵检测系统的通用模型。但是在八十年代,这些理论和模型都没有引起人们的关注。Internet 普及全球之后,入侵检测才真正得到重视,并快速地发展。

随着人类社会步入飞速发展的互联网时代,与此同时黑客和一些网络上的恶意攻击者利用计算机和网络技术频繁进行网络入侵。如今的网络安全技术,包括防火墙技术、VPN 技术、PKI 技术等都是着重于对网络攻击的防护,但是从网络安全发展的趋势来看,做好

收稿日期:2015-01-03 **修回日期**:2015-05-06 **网络出版时间**:2016-01-04
基金项目:国家自然科学青年基金项目(71271071)
作者简介:张公让(1966-),男,副教授,研究方向为商务智能、数值模拟;万 飞(1988-),男,硕士研究生,研究方向为数据挖掘、人工智能。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20160104.1453.004.html

防护的同时,引入入侵检测技术对网络攻击行为进行预测和阻止才是治标又治本的办法。

网络入侵检测一直是人们研究的热点问题,近年来相关人员针对这个问题提出了很多模型和方法,数据挖掘就是其中的一种。具体的、常用的有分类技术、聚类技术、关联规则挖掘等[2]。支持向量机(Support Vector Machine,SVM)便是一种具有良好学习能力和推广能力的分类技术。

SVM 是俄罗斯统计学家和数学家 Vapnic 与其同事于 1995 年首先提出的[3],它在预测小样本、非线性以及高维数据的训练拟合中拥有很多特有的优势。支持向量机方法以统计学习理论里的 VC 维理论和结构风险最小化原理为基础,将训练样本映射到高维空间,并在这个空间里建立最大分类超平面,即最大间隔分类器,如图 1 所示。通过最大间隔分类器对测试样本进行预测。

图 1 使用安德森鸢尾花卉数据集训练得到的最大间隔分类器

基于支持向量机的网络入侵检测方法具有很好的泛化能力,在分类模型建立过程中,核函数的参数 g 和惩罚系数 c 对分类器的性能有很大的影响。但是对于 SVM 参数的优化选取,并没有公认的最好的方法。通常基于 SVM 的网络入侵检测方法的参数选取,一般都采用默认参数或者根据经验设置。实验证明,使用网格搜索法对 SVM 参数进行优化,具有最好的泛化精度。在此基础上,对数据集进行归一化处理,将大幅度减少构建分类器的迭代次数,从而减少训练和预测时间。

目前,SVM 检测技术在文本分类、车辆识别、工业生产等领域都有普遍的应用。张国梁等[4]提出了使用互信息特征选择法结合 GIGMOID 核函数对新闻文本分类的研究,取得了不错的分类准确率;周辰雨等[5]以遗传算法为搜索模式,采用交叉验证技术确定 SVM 的最佳参数组合;贾存良等[6]通过对三种核函数的计算结果对比,选择合适的核函数并应用于煤炭需求预测;张琨等[7]通过实验对比的方式来确定核函数并进行参数选择。上述理论只是单纯比较核函数或者仅采用某一种参数优化算法,预测结果并未达到最优。近年来,王健峰等[8]提出使用改进的网格搜索法对 SVM 参数进行优化,在确定最优参数区间之后,再进行小步距精搜索。该方法适用于样本属性值在同一数量级且无量纲的数据。网络连接数据的属性值变化很大,在参数优化之前,若不采用数据归一化处理,会降低对样本的预测成功率且极大地增加迭代计算量。另外,还有一些学者专注于样本数据集的属性选择。朱文杰等[9]通过信息熵理论对样本数据进行处理,剥离下行属性集,从而降低样本的特征维数,有效减少检测的计算规模;徐永华等[10]提出 K -means 与属性信息熵相结合,对训练样本集进行约简。传统属性约简方法有 Relief[11]算法、K -means 算法、粗糙集理论[12]等。

2 网格搜索算法和数据归一化

2.1 交叉验证

交叉验证(Cross Validation[13],CV)是一种检验分类器泛化能力的统计方法。交叉验证的基本思想是将训练数据切割成 K 个较小的子集,每次迭代,使用一个子集做测试集,其余 K -1 个子集作为训练集进行分析。这种分组方法,被称为 K 折交叉验证(K -fold CV)。K 折交叉验证可以有效选取支持向量机的最优核参数和最优惩罚系数,同时避免出现过度拟合的发生,因此,实验结果具有很强的说服力。实验中采用 10 折交叉验证,将数据集分为 10 份,依次将其中 9 份作为训练数据,另外 1 份作为测试数据进行试验。每次试验都会得到相对应的迭代次数和检测准确率(CVAccuracy)。

2.2 网格搜索法

网格搜索法的基本原理是让 c 和 g 在一定的范围划分网络并遍历网格内所有点进行取值,对于取定的 c 和 g 利用 K -CV 方法得到此组 c 和 g 下训练集验证分类准确率,最终取使得训练集验证分类准确率最高的那组 c 和 g 作为最佳的参数[8]。网格搜索法参数优化的结果如图 2 所示,其中 c 的取值范围设置为[2^{-8},2^{8}],g 的取值范围设置为[2^{-8},2^{8}],参数 c 和 g 的步进大小范围设置为 1。

由图 2 可以看出,c 和 g 在一定的区间上分类准确率较低,造成这一结果的原因是,训练数据中某些列的属性值(如第 5 列)过大,而其他列的属性值相对较小,容易造成训练时某些样本属性值的丢失。因此,对样本数据进行归一化处理是必要的。

2.3 归一化处理

对于每个样本,由于它在每个维度上的量纲不同,如果不对样本进行归一化处理,在量纲数量级差别悬

殊的时候,会使样本中较低数量级的属性变为 0,从而会使原来样本数据的信息丢失过多。

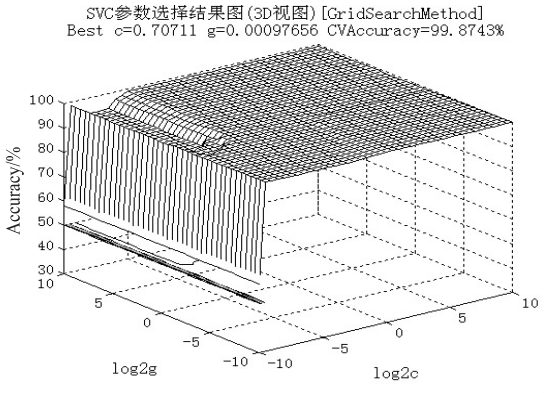

图 2　网格搜索法参数选择结果

2.3.1　Min-Max 标准化(Min-Max Normalization)

标准化是对原样本数据的线性变换,也称为离差标准化。通过变换使样本数据映射到 0 ~ 1 之间。转换函数如下所示:

$$x^* = \frac{x - \min}{\max - \min}$$

2.3.2　训练集和测试集合并归一化

如果现将训练集进行归一化(假设第一维度的最大值为 M),并将这个归一化映射记录下来,当有新的测试集(假设第一维度的最大值为 N)时再用这个归一化映射对测试集进行归一化。这样,就接受了这样一个假设:N 不能超过 M。不然归一化会产生结果大于 1 的不合理情况。但是,将训练数据和测试数据放在一起归一化就可以避免出现这种情况,归一化后每一维度的最大值和最小值是从训练数据和测试数据的集合中寻找。

做这样的处理会出现一个问题,每次变换测试数据,都要对分类器进行重新训练,较为耗费时间。但是,需要考虑到所处理的问题是面向数据的,当加入了新的测试数据时,如果建立一个更加适合这个测试集的 SVM 模型,预测结果将更加准确和合理。归一化处理后的参数寻优效果如图 3 所示。可以看出,当取得最佳参数的同时,分类准确率也有一定提升。

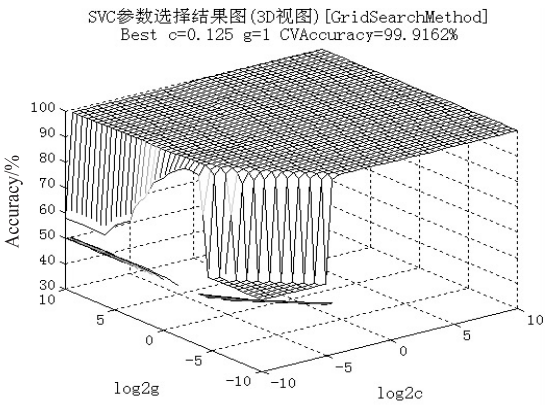

图 3　归一化+网格搜索法参数选择结果

2.4　SVM 入侵检测过程

使用数据归一化和网格搜索法的网络入侵检测过程如图 4 所示。

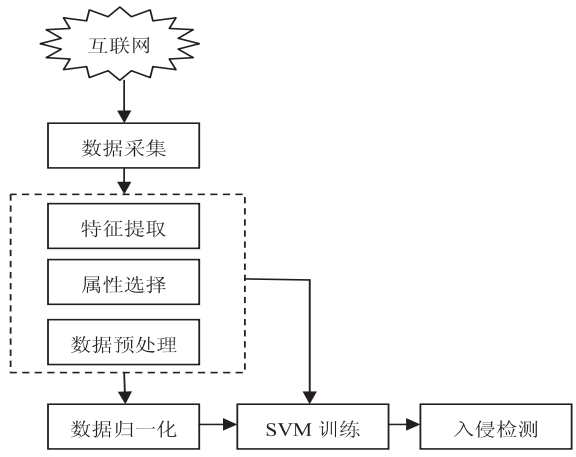

图 4　网络入侵检测过程

具体步骤如下:

步骤 1:在互联网中获取网络中的数据包信息。通常使用网络嗅探器来实现网络数据的获取;

步骤 2:对获取的数据进行简单的特征抽取和属性选择,去除含字符串的三列属性值,并将标签列数据变换成正常与攻击两种类型,方便进行二分类处理;

步骤 3:对属性列进行归一化处理,将数据训练集和测试集合并起来进行归一化处理。集成之后的数据集所建立的 model 更能反映测试集的性质,因而可以获得更高的分类准确率;

步骤 4:使用网格搜索算法对 SVM 进行参数寻优。使用 10 折交叉验证,限定惩罚参数 c 和 RBF 核函数 g 的变化范围都在 $[2^{-8}, 2^{8}]$,限定 c 和 g 的步进大小为 1,步进间隔大小为 4.5;

步骤 5:SVM 核函数选择 RBF 核函数。使用最优参数训练分类器,得到最优参数所对应的 model,将测试数据代入该 model 中计算,得到迭代次数和分类准确率;

步骤 6:分别使用遗传算法和粒子群算法对 SVM 参数进行寻优,将实验结果和步骤 5 中得到的实验结果进行对比。其中遗传算法中的参数终止代数设为 50,种群数量 pop 设为 20;粒子群算法中的参数终止代数设为 100,种群数量 pop 设为 20。实验中将未归一化数据也代入 model 进行计算,得出结果进行对比。

3　实验结果与分析

3.1　实验数据

文中实验数据集采用 KDDCUP 99[14] 数据集。1998 年美国国防部高级规划署(DARPA)在 MIT 林肯实验室进行了一项网络入侵检测评估项目,通过模拟

美国空军局域网,收集了长达 9 周的网络连接和审计数据,仿真各种攻击手段。每个网络连接都被标记为 normal 或 attack,异常类型有四大类:PROBE、DOS、U2R 和 R2L,其中 DOS 攻击最多。异常细分为 39 种攻击类型。实验工具使用台湾大学林智仁教授编写的 LIBSVM 工具和 Matlab 软件。

实验中,从 KDDCUP 99 数据集的训练样本集中分别选取 212、520、2 387 条数据作为候选训练样本数据。从 KDDCUP 数据集的测试集中选取 30 000 条数据作为测试样本数据。

3.2 实验对比

实验分为两步:

1)采用传统 SVM、基于遗传算法(GA)的 SVM 参数优化、基于粒子群算法(PSO)的 SVM 参数优化和基于网格搜索的 SVM 参数优化进行对比。实验结果如表 1 所示。

表 1 参数寻优对比

训练样本数	测试样本数	传统 SVM /%	PSO 寻优/%	GA 寻优/%	网格寻优/%
212	30 000	92.646 7	95.096 7	94.926 7	95.11
520	30 000	93.46	95.103 3	94.926 7	95.143 3
2 387	30 000	94.026 9	95.14	94.926 7	95.166 7

可以看出:

(1)随着训练样本数的提高,分类准确率也相应提高;

(2)三种参数优化算法均有较大幅度提高;

(3)和其他优化算法相比,基于网格搜索的参数寻优分类效果最好。

2)使用数据归一化和未使用数据归一化处理的 SVM 参数优化对比,以及相应建模迭代次数和训练数据拟合程度的对比。实验结果如表 2 ~4 所示。

表 2 归一化前后准确率对比 %

	传统 SVM	PSO 寻优	GA 寻优	网格寻优
未归一化	94.026 9	95.14	94.926 7	95.166 7
归一化	94.99	95.173 3	95.153 3	95.243 3

表 3 归一化前后迭代次数对比

	传统 SVM	PSO 寻优	GA 寻优	网格寻优
未归一化	1 737	1 699	2 141	1 361
归一化	28	40	95	180

表 4 归一化前后 CVAccuracy 的对比 %

	PSO 寻优	GA 寻优	网格寻优
未归一化	99.706 7	98.994 6	99.874 3
归一化	99.916 2	100	99.916 2

可以看出:

(1)数据归一化后,对测试数据的分类准确率有一定的提升;

(2)建立分类器的迭代次数大幅减少,当测试数据很大的时候,会极大地缩短测试和系统应对的时间;

(3)数据归一化之后,分类器对训练数据的拟合度提升,同时可以提升对测试数据的分类准确率,说明并未过拟合。

4 结束语

文中探讨了将网格搜索技术和数据归一化方法应用于基于 SVM 的网络入侵检测的系统中,以解决传统 SVM 技术在检测时间和分类准确率方面的问题。结果表明,通过对样本数据进行归一化处理,并采用网格搜索技术对 SVM 的参数进行优化,可以减少检测网络异常所需的时间并提高检测的准确率,是网络入侵检测算法优化中一次有效的尝试。

参考文献:

[1] 雷渭侣,王玉兰. 计算机网络安全技术与应用[M]. 北京:清华大学出版社,2009.

[2] 倪志伟,倪丽萍,刘惠婷,等. 动态数据挖掘[M]. 北京:科学出版社,2010.

[3] Cortes C, Vapnic V. Support-vector networks[J]. Machine Learning,1995,20(3):273-297.

[4] 张国梁,肖超锋. 基于 SVM 新闻文本分类的研究[J]. 电子技术,2011,38(8):16-17.

[5] 周辰雨,张亚岐,李 健. 基于 SVM 的车辆识别技术[J]. 科技导报,2012,30(30):53-57.

[6] 贾存良,吴海山,巩敦卫. 煤炭需求量预测的支持向量机模型[J]. 中国矿业大学学报,2007,36(1):107-110.

[7] 张 琨,曹宏鑫,严 悍,等. 支持向量机在网络异常入侵检测中的应用[J]. 计算机应用研究,2006,23(5):98-100.

[8] 王健峰,张 磊,陈国兴,等. 基于改进的网格搜索法的 SVM 参数优化[J]. 应用科技,2012,39(3):28-31.

[9] 朱文杰,王 强,翟献军. 基于信息熵的 SVM 入侵检测技术[J]. 计算机工程与科学,2013,35(6):47-51.

[10] 徐永华,李广水. 基于距离加权模板约简和属性信息熵的增量 SVM 入侵检测算法[J]. 计算机科学,2012,39(12):76-78.

[11] Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection[C]//Proc of international joint conference on artificial intelligence. [s. l.]:[s. n.],2001.

[12] 张文修. 粗糙集理论与方法[M]. 北京:科学出版社,2001.

[13] Kononenko I. Estimating attributes:analysis and extensions of relief[C]//Proc of ECML. [s. l.]:[s. n.],1994:171-182.

[14] Kdd B. KDD99 cup dataset[EB/OL]. 1999. http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html.

基于网格搜索的 SVM 在入侵检测中的应用

作者： 张公让， 万飞， ZHANG Gong-rang， WAN Fei
作者单位： 合肥工业大学 管理学院,安徽 合肥,230009
刊名： 计算机技术与发展 ISTIC
英文刊名：
年，卷(期)： 2016(1)