DOI:CNKI:50-1165/N.20120915.2319.015

一种改进的 SVM 多类分类算法在入侵检测中的应用*

李太白，唐万梅
(重庆师范大学 计算机与信息科学学院，重庆 401331)

摘要:入侵检测作为网络安全的关键技术,成为了当前网络安全研究的热点,入侵检测算法的准确率和推广性能是研究的重点。基于二叉树的思想和超球支持向量机的特点,本文提出了一种改进的 SVM 多类分类入侵检测算法。本文通过引入相似度函数作为权值,选取相似性最小的两类样本构造两类分类器,采用自下而上的方法构造多个两类超球 SVM 分类器,并将该多类分类算法应用于入侵检测中。利用 KDD CUP 1999 入侵检测数据进行了仿真实验,实验结果表明,该算法能有效提高检测准确率、推广性能也得到较好改善。

关键词:支持向量机; 球结构; 二叉树; 入侵检测

中图分类号:TP391 **文献标志码**:A **文章编号**:1672-6693(2012)05-0063-04

支持向量机(Support vector machine,SVM)[1]是在统计学习理论基础上提出的一种解决小样本、高维、非线性等问题的有效方法。SVM 是针对两类分类问题提出的,而实际问题大多是多类分类问题。目前对于多类分类问题[2],主要有一对一方法、一对多方法、DAGSVM 方法、二叉树支持向量机等。1999 年 Tax 和 Duin 提出了生成超球的方法[3],在此基础上 Zhu 与 Wang 等提出支持向量机多类分类算法[4]。

入侵检测作为网络安全的关键技术,成为了当前网络安全研究的热点。目前已有统计方法[5]、神经网络方法[6-7]、支持向量机方法[8-9]等,这些方法有各自的优缺点。统计方法依赖于一些假设,如审计数据或用户行为的分布符合高斯分布。实际上,用户行为具有随机性,这种假设可能导致较高的误警率。对于神经网络在入侵检测中的应用,大多数研究工作中都是基于 BP 网络实现的,但是,BP 网络在实验中暴露出很多性能上的不足,诸如在学习时陷入局部极小、学习时间长、不易收敛等。也有人提出了基于 SVM 的入侵检测算法,虽然在训练时间和学习泛化能力上都有一定的改善,但精度不高,并不能很好地区分异常数据。

针对以上问题,本文提出了一种改进的 SVM 的多类分类入侵检测算法。根据二叉树的思想和球结构 SVM 的特点,该算法采用相似度函数作为权值构造多个两类超球 SVM 分类器,用于解决多类分类问题,并将该算法应用于入侵检测中以提高系统性能。

1 超球支持向量机理论

给定 N 类问题的集合 A^m,$m=1,\cdots,N$,每个集合包含 l^m 个样本点 x_i^m,$i=1,\cdots,l^m$,对每个集合寻找一个超球(a^m,R^m),a^m 为球心,R^m 为半径平方(x_i^m 尽可能地达到最小),使得最小超球尽可能包含所有的同类样本点 x_i^m。初始的最优化问题为

$$\min R^m + C^m \sum_i^{l^m} \xi_i^m \tag{1}$$

$$\text{s.t. } \| \chi_i^m - a^m \|^2 \leqslant R^m + \xi_i^m, \xi_i^m \geqslant 0,\ i=1,\cdots,l^m$$

式中,ξ_i^m 为松弛变量;C^m 为惩罚系数。将非线性分类的当前训练集,通过一个非线性映射 $\phi(x_i^m)$,把训练数据 x_i^m 映射到高维线性特征空间。但在求解时不需要计算该非线性函数,以二分类为例,只需计算核函数 $K(\chi_i,\chi_j)=\langle\phi(\chi_i),\phi(\chi_j)\rangle$。采用拉格朗日乘子法求解这个具有线性约束的二次规划问题,得到的对偶优化问题为

* **收稿日期**:2011-12-05 **修回日期**:2012-02-27 **网络出版时间**:2012-9-15 23:19
资助项目:重庆市教委科学技术项目(No. KJ110617);重庆市自然科学基金(CSTC2010BB2090);重庆师范大学校级项目(No. cyjg1205)
作者简介:李太白,男,硕士研究生,研究方向为智能计算;**通讯作者**:唐万梅,E-mail:cqtwm@163.com
网络出版地址:http://www.cnki.net/kcms/detail/50.1165.N.20120915.2319.201205.63_015.html

$$\min \quad Q(\alpha)=\sum_{i=1,j=1}^{n}\alpha_i\alpha_j K(\chi_i,\chi_j)-\sum_{i=1}^{n}\alpha_i K(\chi_i,\chi_i) \tag{2}$$

$$\text{s.t.}\ \sum_{i=1}^{n}\alpha_i=1,\ 0\leqslant\alpha_i\leqslant C$$

式中 x 为已知量,唯一的变量是 a,球心可表示为

$$a=\sum_{i=1}^{n}\alpha_i\phi(\chi_i) \tag{3}$$

超球半径 R 可由在超球内的点确定

$$R^2=\|\phi(\chi)-a\|^2=K(\chi,\chi)-2\sum_{i=1}^{m}\alpha_i K(\chi,\chi_i)+\sum_{i,j=1}^{m}\alpha_i\alpha_j K(\chi_i,\chi_j) \tag{4}$$

在数据集中的任意一点 x_i 到超球心 a 的距离为

$$d(x_i)=\|\phi(\chi_i)-a\|=\sqrt{K(\chi_i,\chi_j)-2\sum_{j=1}^{m}\alpha_j K(\chi_i,\chi_j)+\sum_{j,k=1}^{m}\alpha_j\alpha_k K(\chi_j,\chi_k)} \tag{5}$$

2 基于二叉树的球结构 SVM 多类分类算法

二叉树法(Binary tree, BT)的思想[10]是先将所有类别划分为两个子类,每个子类又划分为两个子类。该方法将原有的多类分类问题分解为一系列的两类分类问题,其中两个子类间的分类函数采用SVM进行训练。文献[11]认为,越是上层节点的SVM子分类器的分类性能对整个模型的推广性影响越大。因此,在生成二叉树的过程中,应该让最易分割的类最早分割出来,即在二叉树的上层节点处分割。

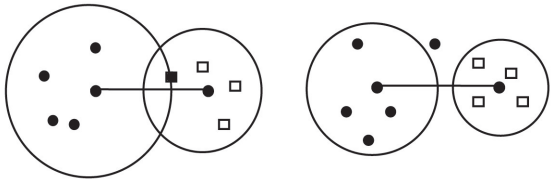
基于此思想,本文采用自下而上的方法构造二叉树,即先寻找最难分割的两类作为二叉树的下层结点,然后再寻找次难分割的两类,直到只包含一个类别为止。另外,球结构 SVM 对不平衡样本有很好的处理能力。基于此,本文结合上述两者提出了一种改进的多类分类算法(SSVM)。

2.1 相似度函数

欧氏距离被广泛用于相似性的度量,但两个类的中心距离并不能准确地反映出两类的相似度。如图1所示的两类,它们的类中心距离相同,但两类的位置关系却不同。其中,图1(a)为两类相交;图1(b)为两类相离。显然图1(a)比图1(b)具有更高的相似性。因此,不能只以类中心距离远近作为相似性度量的标准,还应该考虑类内样本的分布情况。球结构的SVM能构造出使半径最小且尽可能包含该类所有样本的球体,因此球体的半径可以用来度量类内样本的分布。由上分析,用下面的距离计算方法[12]作为类 i 与 j 间相似性度量

$$d_{ij}=\|(a^i-a^j)\|_2-(R^i+R^j) \tag{6}$$

其中:a^i,a^j,R^i,R^j 分别是第 i 类和第 j 类的中心和半径。当 $d_{ij}\geqslant 0$ 时表明第 i 类与第 j 类没有相交区域。d_{ij} 的值越大说明第 i 类与第 j 类的可分性越强;d_{ij} 的值越小说明两类越相似。



(a) 2个球体相重叠 (b) 2个球体相分离

图1 2个球体的分布

2.2 算法描述

在对 N 类训练样本进行训练生成二叉树结构时,首先是将相似性 d_{ij} 最小的两类合并为 s_1 作为其父结点,同时训练生成SVM子分类器。然后删除已被选取的两类样本并分为在集合 S 中有 s_1 和没有 s_1 的情况进行考虑。从 S 中选取相似性最小的两类样本合并为 s_2,训练另一个SVM子分类器。在计算某类与 s_1 的相似性时,采用到 s_1 中各样本的距离和的均值作为相似性度量标准。直至最终得到二叉树的根结点,即训练完成。对待分类样本的分类过程,为训练过程的逆方向,从根结点开始,依次序经过若干个两类分类器到达某个叶子结点,该结点对应的类即为该样本的类别。

具体步骤如下:

1) 将 N 类样本放入集合 S 中,其中,N 是样本的类别数。

2) 根据距离公式(6),在集合 S 中选取 d_{ij} 最小的两个类 A_1,A_2。将半径较小的一类样本作为正类,另一类作为负类。即if $R_{A_1}<R_{A_2}$,则将 A_1 作为正类,A_2 作为负类;进行训练得到SVM子分类器SVM_1,将这两类合并为 s_1。

3) 在集合 S 中删除被选取的样本,即在 S 中删除 A_1,A_2 的样本。if $|S|=\phi\ \|\ |S|=1$,算法结束;否则:

① 根据距离公式(6)在集合 S 中选取 d_{ij} 最小的两类样本 B_1，B_2，距离值标记为 d'_{ij}；

② 将集合 s_1 加入 S 中。根据距离公式(6)在集合 S 中选取某类 C，使其与集合 s_1 中各样本的距离和的均值最小，将距离和的均值标记为 d_{avg}。

4) 根据 d'_{ij}，d_{avg} 进行比较。

① if $d'_{ij}<d_{avg}$，选取 B_1，B_2，同理进行训练得到 SVM 分类器 SVM_2，将这两类合并得到集合 s_2。

② else if $d'_{ij}>d_{avg}$，选取 C，s_1 进行训练得到 SVM 分类器 SVM_2，将这两类合并得到集合 s_2。

5) 重复步骤 3 和步骤 4，直到 S 中仅剩一棵二叉树为止，这棵二叉树即为所求。

3 实验验证及比较分析

3.1 数据集描述

本文采用的实验数据来源于 KDD CUP 1999 数据集，为了验证本文提出的基于球结构的 SVM 多类分类算法的准确率，以及多类 SVM 用于入侵检测系统的有效性，作者从 KDD CUP 1999 原始数据集中选择了具有代表性的 4 000 多个数据，数据集中每条记录包含 41 维特征，根据美国 DARPA 的入侵检测评估报告，目前常用的攻击手段和方法分为以下 4 种：1)DOS，拒绝服务攻击；2)R2L，远程用户到本地的非授权访问；3)U2R，非授权获得超级用户权限攻击；4) Probe，探测攻击。在进行检测之前，将经过预处理后的数据集随机分成一个训练集(Train 60%)和一个测试集(Test 40%)，实验是在 Normal、DOS、Probe、U2R、R2L 5 个类别之间进行的。同时该算法(SSVM)与标准支持向量机(SVM)和文献[13](PSVM)的入侵检测算法进行了比较。实验所采用的数据集的特性如表 1 所示。

表 1　入侵检测的实验数据集

序号	数据集	训练集	测试集	类别名
1	1 150	690	460	Normal
2	1 050	630	420	Dos
3	1 350	810	540	Probe
4	100	60	40	U2R
5	1 200	720	480	R2L

3.2 实验评价标准和结果分析

入侵检测系统的性能评价指标主要有 3 个，它们分别是检测率、误报率、准确率，这些指标的定义如下。

检测率(Detection rate，DR)＝被检测出的攻击样本数/异常样本的总数。

误报率(False positive rate，FPR)＝正常样本被误报为异常样本数/正常样本数。

准确率(Accuracy rate，AR)＝各类被正确分类的样本数总和/各类样本数总和。

实验结果表明，在 DOS 类数据上，SSVM 与 PSVM 相比，有相同的检测率，误检率稍低；在 U2R 类数据上，SSVM 与 PSVM 相比，检测率明显高于 PSVM，虽然误检率比 PSVM 稍高，但在该数据上的综合正确率要高于 PSVM；在 R2L 数据上，SSVM 与 PSVM 相比，检测率稍高于 PSVM，误检率较低；在 Probe 数据上，SSVM 与 PSVM 相比，检测率稍高于 PSVM，误检率低于 PSVM。显然，采用基于球结构 SVM 的多分类算法应用于入侵检测中取得较为理想的结果，实验结果表明了该算法在准确率、推广性能方面得到有效的提高。表 2 中，黑体部分表示在对应指标上取得相对较好的结果。

表 2　在入侵检测数据上的实验　　%

算法	攻击类别				准确率
	DOS	Probe	U2R	R2L	
SVM[1]	$DR=99.32$ $FPR=0.91$	$DR=97.26$ $FPR=0.52$	$DR=69.11$ $FPR=0.58$	$DR=98.11$ $FPR=1.99$	95.20
PSVM[13]	$DR=100$ $FPR=0.74$	$DR=98.97$ $FPR=0.45$	$DR=68.97$ $FPR=0.15$	$DR=98.61$ $FPR=1.58$	98.36
SSVM	$DR=100$ $FPR=0.53$	$DR=99.16$ $FPR=0.37$	$DR=73.71$ $FPR=0.42$	$DR=98.84$ $FPR=1.23$	98.76

4 结束语

入侵检测作为网络安全的关键技术,成为当前网络安全研究的热点,入侵检测算法的准确率和推广性能是研究的重点。基于二叉树的思想和超球支持向量机的特点,本文提出了一种改进的 SVM 多分类入侵检测算法。引入了相似度函数作为权值,选取相似性最小的两类样本构造两类分类器,采用自下而上的方法构造多个两类超球 SVM 分类器,并将该多类分类算法应用于入侵检测中。在 KDD CUP 1999 数据集上进行了仿真实验,实验结果表明该算法能有效提高准确率、并改善了推广性能。在此基础上,结合其他方面的相关技术选择一种结构更简洁、计算更有效率的 SVM 多类分类算法,建立正确率高、误报率低、检测速度快的入侵检测系统服务是进一步的研究方向。

参考文献:

[1] Vapnik V. The nature of statistical learning theory [M]. New York:Springer,1995.

[2] Hsu C W,Lin C J. A comparison of methods for multi-class support vector machines[J]. IEEE Transactions on Neural Networks, 2002,13 (2):415-425.

[3]Tax D M J,Duin R P W. Data domain description using support vectors [C]//Anon Poceedings of European Symposium on Artificial Neural Networks. Bruges (Belgium): D-Facto,1999:251- 256.

[4] Zhu M L,Wang Y, Chen S F, et al. Sphere-structured support vector machines for multi-class pattern recognition [J]. Lecture Notes in Computer Science, 2003, 2639:589-593.

[5] Bykova M, Ostermann S, Tjaden B. Detecting network intrusions via a statistical analysis of network packet characteristics [C]//Anon Proceedings of the 33rd Southeastern Symposium on System Theory. Athens: IEEE,2001:309-314.

[6] Han S J,Cho S B. Evolutionary neural networks for anomaly detection based on the behavior of a program [J]. IEEE Transactions on Systems,Man,and Cybemetics,Part B,2005:559-570.

[7] 黄勤,龚海清,刘金亨. 基于改进的遗传神经网络入侵检测系统[J]. 重庆理工大学学报:自然科学版,2010,24(2):83-86.

[8] Cao L J,Chua K S,Chong W K. A comparision of PCA, KPCA and ICA for dimensionality reduction in Support Vector Machine[J]. Neurocomputing,2003,55(2):321-336.

[9] 张晨,王晓东. 基于支持向量机的网络入侵异常检测[J]. 重庆工学院学报:自然科学版,2007,21(12):119-121.

[10] 秦玉平,罗倩,王秀坤,等. 一种快速的支持向量机多类分类算法[J]. 计算机科学,2010,37(7):240-242.

[11] 刘洋,张秋余. 基于 huffman 树的多类 SVM 方法[J]. 计算机工程与设计,2008,29(7):1792-1840.

[12] 谢志强,高丽,杨静. 改进的球结构 SVM 多分类增量学习算法[J]. 哈尔滨工程大学学报,2009,30(9): 1041-1046.

[13] 顾钧. 基于 KPCA 和 SVM 的网络入侵检测研究[J]. 计算机仿真,2010,27(7):105-107.

Application of an Improved SVM Multi-Class Classification to Intrusion Detection

LI Tai-bai, TANG Wan-mei

(College of Computer and Information Science, Chongqing Normal University, Chongqing 401331, China)

Abstract: Intrusion detection system as the key technology of network security becomes research hot spot of the current network security, while precision and generalization performance is the key point of intrusion detection algorithm. According to binary tree method and the characteristics of sphere structured support vector machine, an improved SVM multi-class classification algorithm is proposed to intrusion detection. This algorithm uses similarity functions as weight value and selects two kinds of sample similarity minimum to structure two-class classifier; to bottom-up structure kinds of two-class classifier of sphere structured SVM. Finally it is applied to intrusion detection. The KDD CUP 1999 intrusion detection data used to simulate experiments. Experimental results show that the algorithm effectively improved the detection accuracy and generalization performance.

Key words: Support Vector Machine; sphere structure; binary tree; intrusion detection

(责任编辑 游中胜)