基于稀疏编码和 SVM 的协同入侵检测*

崔 振[1,2],陈柏生[1]
(1.华侨大学 计算机科学与技术学院,福建 厦门 361021;
2.中国科学院 计算技术研究所,北京 100190)

摘 要:将稀疏编码理论应用于入侵检测,并提出一种将稀疏编码理论和支持向量机结合的入侵检测算法。稀疏性约束同时引入到过完备词典学习和编码过程,学习到的系数作为特征送入到支持向量机进行入侵检测。实验表明,稀疏性具有一定的去噪能力,使得学习的特征更富有判别力。同时实验也验证了所提出的方法能保证较高的检测率和较低的误报率,并且对不平衡数据集有较好的鲁棒性。

关键词:稀疏编码;支持向量机;协同;入侵检测;过完备词典
中图分类号:TP181 **文献标识码**:A **文章编号**:1674-7720(2011)22-0078-04

Cooperative intrusion detection system based on sparse coding and support vector machine

Cui Zhen[1,2], Chen Baisheng[1]
(1.College of Computer Science & Technology, Huaqiao University, Xiamen 361021, China;
2.Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: The theory of sparse representation is applied to intrusion detection, and an approach based on sparse coding and support vector machine is also proposed for intrusion detection. Sparsity constraints are added to train the over-complete dictionary and encode samples simultaneously. Learned sparse coefficients as features are fed into support vector machine for intrusion detection. Experiments show that the sparsity can remove some noises and make mapping features more discriminative. Meanwhile, experiments also prove our proposed method more effective with higher detection rate and lower false alarm rate, especially good robustness in the imbalanced dataset experiment.

Key words: sparse coding; SVM; cooperation; intrusion detection; over-complete dictionary

将所有的网络行为分成正常行为和异常行为两类,这样入侵检测问题就可以转化成模式分类问题。入侵检测的关键是正常和异常行为模式库的建立。目前常用的入侵检测方法有基于贝叶斯推理的入侵检测[1]、基于模式匹配的入侵检测[2]、基于神经网络的入侵检测[3]和基于数据挖掘的入侵检测[4],以上方法对数据的要求较高或需要的数据量较大。支持向量机是建立在统计学习理论上的一种新的机器学习方法,由于其在小样本、高维、非线性等方面的优势和较好的推广能力,已经在入侵检测中得到应用[5]。总体上,支持向量机可分为误用检测和异常检测两大类,误用检测准确度高,但难以应对未知攻击;异常检测则常常面临误报率过高的问题。另外,如何应对大规模的高速数据流检测、如何实现在线学习、如何减少或消除噪声数据的影响,是入侵检测系统面临的主要挑战。

近年来,稀疏表示相关理论已成为研究的热点。常用的信号分解方式通常是非冗余的正交变换,如离散余弦变换、小波变换等。这类方式缺乏灵活性,并且许多混合信号在单一的正交基变换中无法得到有效的稀疏表示。基于超完备字典的信号稀疏分解是一种新的信号表示理论,它采用冗余原子来构造字典,而不是采用传统的正交基,这样使字典更富有表现力,同时为信号自适应的稀疏扩展提供了空间。通过这种超完备字典把数据变换到另一空间,即进行稀疏编码,会带来更好的分类效

*基金项目:国务院侨办科研基金资助项目(10QZR0);华侨大学科研基金资助项目(10HZR06)

果[6]，原因是稀疏表示系数从某种意义上带有一定的判别信息[7]。稀疏表示已应用于一些具体的领域：学习非参数化字典来进行图像超分辨率或图像重建[8]；利用稀疏表示系数重构图像，用重构误差进行(遮挡)人脸识别[7]等。这些应用领域主要集中在图像处理和压缩感知中。

本文将稀疏编码方法应用于入侵检测。在过完备词典学习和编码的过程中加入 l_1 范数约束，同时最小化重构残差和非零个数，在去除一定噪声的同时也促使映射的特征本身具有稀疏性。这种稀疏性使得学习到的系数特征拥有更好的判别性，即学习后的特征在分类空间更易于划分，同时后端结合强大的分类器——支持向量机来进行入侵检测。实验中，本文所提的方法与直接用 SVM 的方法进行了比较，显示了稀疏映射的特征更富有表示力和判别力，验证了所提方法的有效性。

1 稀疏表示理论

1.1 词典学习

构建字典归纳起来有两种方法[9]：(1)基于数据模型建立稀疏字典，如一些小波函数；(2)从训练集中学习一个字典。本文采用后一种方法构建字典。由于数据的规模很大，采用较流行的字典训练方法——SVD 分解来迭代构建词典，即 K-SVD[10]方法。

K-SVD 算法由 K-均值聚类算法推广而来，是一种迭代方法，一方面用当前字典对训练集信号进行稀疏编码，另一方面更新字典的原子以期使得字典更好地表示信号。这种联合的更新加速了算法的收敛。K-SVD 算法是灵活的，可以和任何一种追踪算法一起工作。K-SVD 的目标函数：

$$\min_{D,X} \|Y-DX\|_F^2 \quad \text{s.t.} \quad \forall i, \|x_i\|_0 \leqslant T_0 \tag{1}$$

其中，$Y=(y_1, y_2, \cdots, y_N)$，$y_i \in R^n$ 是第 i 个样本，$D=[d_1, d_2, \cdots, d_k] \in R^{n\times K}$ 是词典，K 是原子数量，X 是稀疏系数，$\|\cdot\|_0$ 是 l0 范数。

K-SVD 算法分为两步：

(1)固定 D，更新稀疏系数 X。可以通过任何稀疏编码算法求解，如 LARS、OMP、BP 等。

(2)同时更新 D 和 X。采用 SVD 分解用最大奇异值对应的特征向量来更新字典。记字典 D 第 k 列为 d_k，对应的稀疏系数为 x_R^i(X 的第 i 行)，式(1)可表示为：

$$\|Y-DX\|_F^2 = \left\| \left[Y - \sum_{i\neq k} -d_i x_R^i \right] - d_k x_R^k \right\|_F^2 = \left\| E_k - d_k x_R^k \right\|_F^2 \tag{2}$$

然后对 E_k 应用 SVD 分解：$E_k=U\triangle V$。令 $d_k=U(:,1)$，$x_R^i=\triangle(1,1)\times V(:,1)^{\mathrm{T}}$。

重复上述两步到规定的迭代次数为止。

1.2 稀疏求解

给定超完备字典 $D \in R^{n\times K}$，其中 $n<K$。测试样本 $y \in R^n$，把测试样本 y 表示成字典原子项 $\{d_i\}(i=1, \cdots, m)$ 的稀疏线性组合，将目标形式化为如下的目标函数：

$$l_1: \hat{x}_1=\arg\ \min\|x\|_1 \quad \text{s.t.}\|Dx-y\|_2<\varepsilon \tag{3}$$

式(3)可在多项式时间内求解。

目前，求解超完备稀疏表示最优化问题的稀疏优化方法主要有贪婪算法、全局优化算法以及其他算法[11]。贪婪算法通过选取字典中与信号最匹配的项，迭代地构造出信号的逼近。全局优化方法是指在满足一定的优化条件下，使得某个特殊的目标函数最小，典型的目标函数是凸函数，并且任何局部最小值也是全局最小值。

本文使用的是 Efron 等提出的 LARS 变量选取方法[12]。算法大致描述如下：

首先稀疏系数设置为 0。然后在词典里查找与响应变量相关最大的输入变量，在响应变量的投影方向选取最大的步长，使得其余的某一个输入变量与当前的输入变量有同样的相关性(在当前的重构残差情况下)。这时候选取了两个变量，由这两个变量组成一个子空间，重构残差在子空间上的投影方向继续前进直到第三个变量进入最相关的集合。这样持续下去直到设定的阈值为止。

LARS 计算的好处是 LARS 路径逐点线性，LARS 的目标函数值是逐步下降的。

2 算法流程

至此，给出基于稀疏编码和 SVM(简记为 SR_SVM)的入侵检测算法流程：

(1)数据预处理

首先把符号类型数值化，然后用下式标准化：$Z_{ji}=(x_{ji}-m(x_i))/\sigma(x_i)$。其中，$m(x_i)$ 表示第 i 个属性的平均值，$\sigma(x_i)$ 为第 i 个属性的标准差，x_{ji} 表示第 j 条记录的第 i 个属性，Z_{ji} 为标准化后的属性值。然后计算标准化度量值，最后把每条记录对应向量单位化，以便于训练字典。

(2)训练字典

设训练集为 Train，对 Train 用 K-SVD 算法[10]训练，超完备字典为 D，$D \in R^{n\times K}$，m 为数据记录维数，n 为词典原子项数。

(3)对训练集求解稀疏表示

对集合 Train 中每个输入的训练样本 y，$y \in R^n$ 使用 LARS 算法[12]最小化 l_1 范数，求解 y 相应于 D 的稀疏表示 $x \in R^K$，并加入到集合 Train_SR。

(4)构建支持向量机模型

使用集合 Train_SR 的数据训练，构建支持向量机模型用于分类检测。

(5)决策分类

设测试集为 Test，对每个测试样本 $y \in R^n$ 使用式(3)最小化 l_1 范数，求解 y 相应于 D 的稀疏表示 $x \in R^K$。使用多类支持向量机对 x 决策类别作为测试样本 y 的类别。

3 实验与分析

实验采用入侵检测领域共同认可及广泛使用的基

准评测数据集——KDD Cup 1999 进行测试。

3.1 实验数据及预处理

实验中使用的训练集和测试集分别从 KDD99 数据集10%的训练子集和测试子集中抽取。为了检验分类器模型的泛化能力，训练集包含 22 种攻击，测试集包含 39 种攻击，训练集中未出现的 17 种攻击占到整个测试集的 10%左右。

KDD Cup 1999 中涉及 3 种协议的数据，分别是 TCP、UDP 和 ICMP。为了更精确地构建冗余字典，加快训练速度，实现并行检测，构建 3 个检测代理，分别是 TCP 检测代理、UDP 检测代理和 ICMP 检测代理(在实际应用中可能拥有更多种类的数据流，可以进行扩展)。根据网络数据流的特点，可以把待检测数据流进行分类(下面分为 3 类：TCP、UDP 和 ICMP，在实际应用中可以扩展)，这样做的前提是假设一次入侵行为不会使用多种网络协议进行通信[13]。针对不同的网络协议，经数据预处理后，就可以去掉一些冗余的属性(在某协议下有些属性的取值是完全相同的)。最后 TCP 选取了 37 个属性，UDP 选取了 20 个，ICMP 选取了 16 个。

3.2 对比实验

实验参数设置：TCP、UDP 和 ICMP 字典的原子项数分别为 60、40 和 40；K-SVD 算法迭代 20 次，稀疏比率约为 10%；SVM 采用 RBF 核函数。

3.2.1 协同检测实验

训练集和测试集抽取情况如表 1 所示。

表 1 数据集抽取情况

	TCP	UDP	ICMP
训练 N 记录数	7 861	9 588	1 288
训练 A 记录数	8 059	1 177	12 219
测试 N 记录数	4 411	1 609	378
测试 A 记录数	9 066	485	6 811

注：N 代表正常数据集，A 代表攻击数据集。

为了说明算法的有效性，将 SR_SVM 与 SVM 进行了对比。实验结果如表 2 所示，可以看到，基于稀疏表示的入侵检测对三种代理都有较高的检测率和较低的误报率。

表 2 协同检测实验结果

	SVM		SR_SVM	
	DR/%	FR/%	DR/%	FR/%
TCP	85.32	0.43	92.86	0.45
UDP	62.47	1.24	88.45	1.43
ICMP	99.87	12.17	99.72	3.97

注：DR 表示检测率(detection rate)，
FR 表示误报率(false alarm rate)。

另外一个值得注意的现象是 UDP 数据集和 ICMP 数据集属于严重不平衡数据集。对于支持向量机来说，这种情况会影响支持向量机超平面的建立。而 SR_SVM 对于不平衡数据集有较好的鲁棒性。

3.2.2 不平衡数据实验

为了进一步测试 SR_SVM 的鲁棒性，在 TCP 数据集上进行不平衡数据集的测验。

测试集不变，继续使用表 1 中对于 TCP 抽取的数据集，训练集分 6 种情况随机抽取，如表 3 所示。实验结果见表 4。从表 4 可以看到，当数据失衡后，相比于数据平衡的情况，检测率有了较大程度的下降，但误报率波动很小，这可能是因为不平衡数据集影响了支持向量机超平面的建立。从结果可以看出，SR_SVM 方法减弱了不平衡数据集对 SVM 的影响，SR_SVM 的检测率较 SVM 有较大程度的提高，误报率基本上与 SVM 持平。无论是正常记录多于攻击记录还是相反情况，SR_SVM 在检测率上基本平稳，而 SVM 的表现则明显差了很多。

表 3 TCP 不平衡记录抽取情况

	训练 N 记录数	训练 A 记录数
训练集 1(4:1)	25 604	6 097
训练集 2(6:1)	19 203	3 157
训练集 3(12:1)	38 406	3 157
训练集 4(1:2)	7 681	16 571
训练集 5(1:5)	7 681	36 990
训练集 6(1:11)	5 120	54 858

注：最左边一列括号里显示的是训练集 N 和 A 的比例。

表 4 不平衡数据实验结果

	SVM		SR_SVM	
	DR/%	FR/%	DR/%	FR/%
训练集 1	68.88	0.29	79.85	0.27
训练集 2	68.16	0.29	76.42	0.29
训练集 3	68.05	0.29	77.27	0.34
训练集 4	76.87	0.36	79.38	0.40
训练集 5	77.47	0.36	82.10	0.41
训练集 6	78.97	0.48	81.78	0.39

3.3 讨论

在分类前，用稀疏编码方法自动提取稀疏特征，而稀疏性符合人类的视觉机理[7]。稀疏性带来好的性能，这在许多文献中已有所体现[7,8]。分析原因主要有两点：

(1)从重构的角度来看，目标函数的第一项是重构残差，最小化重构残差使得系数几乎与原来的样本具有相同的表示能力。

(2)从稀疏的角度来看，使得在保证重构能力的条件下编码稀疏尽量稀疏，即对一些原子具有敏感性，这符合人类的视觉机理[7]。另外，稀疏性起到部分去噪作用，这在大量的图像修复文献中已得到证实[8,14]。因此，稀疏性促使很强的判别力。

本文将稀疏编码与多类支持向量机结合应用到网络入侵中的数据分类，初步的实验结果已显示稀疏性所带来的好处。学习得到的过完备词典可以丰富地表示所有的样本，在词典上稀疏编码也可以有效地学习到样本的判别力特征。在接下来的实验中，会加入更多的实时

数据来完善系统,构建可以应用到实际的实时高效的入侵检测系统。

参考文献

[1] 焦从信,王崇骏,陈世福.基于完全无向图的贝叶斯分类器在入侵检测中的应用[J].计算机科学,2008,35(9):83-86.

[2] 姜庆民,吴宁,刘伟华.面向入侵检测系统的模式匹配算法研究[J].西安交通大学学报,2009,43(2):58-62.

[3] 刘衍珩,田大新,余雪岗,等.基于分布式学习的大规模网络入侵检测算法[J].软件学报,2008,19(4):993-1003.

[4] 刘在强,林东岱,冯登国.一种用于网络取证分析的模糊决策树推理方法[J].软件学报,2007,18(10):2635-2644.

[5] CHEN R C,CHEN S P.An intrusion detection based on support vector machines with a voting weight schema [J]. IEA/AIE 2007:1148-1157.

[6] YANG J,YU K,HUANG T.Efficient highly over-complete sparse coding using a mixture model[C].The 11th European Conference on Computer Vision(ECCV),Crete,2010.

[7] WRIGHT J,YANG A,GANESH A,et al.Robust face recognition via sparse representation[J].IEEE Transactions on Pattern Analysis and Machine Intelligence(PAMI),2009,31(2):210-227.

[8] YANG J,YU K,HUANG T,et al.Image super-resolution as sparse representation of raw image patches[C].In:IEEE Conference on Computer Vision and Pattern Recognition,(2008),Anchorage,AK.

[9] RUBINSTEIN R,BRUCKSTEIN A M,ELAD M.Dictionaries for dparse tepresentation modeling[C].Proceedings of the IEEE,2010,98(6).

[10] Aharon M,ELAD M,BRUCKSTEIN A M.The K-SVD:an algorithm for fesigning of overcomplete dictionaries for sparse representation[J].IEEE Trans.on Signal Processing,2006,54(11):4311-4322.

[11] ZIBULEVSKY M,ELAD M.L1-L2 optimization in signal and image processing[J].IEEE Signal Processing Magazine,2010,27(3):78-88.

[12] EFRON B,JOHNSTONE I,HASTIE T,et al.Least angle regression[J].Ann.Statist,2004,32(2):407-499.

[13] TENG S H,DU H L,WU N Q,et al.A cooperative network intrusion detection based on fuzzy SVMs[J].Journal of Network,2010,5(4):475-483.

[14] MAIRAL J,BACH F,PONCE J,et al.Non-local sparse models for image restoration[C].International Conference on Computer Vision,Tokyo,Japan,2009.

(收稿日期:2011-07-28)

作者简介:

崔振,男,1981年生,讲师,博士研究生,主要研究方向:模式识别,图像处理,数据挖掘。

陈柏生,男,1980年生,讲师,硕士,主要研究方向:模式识别,图像处理。