第37卷 第21期
Vol.37 No.21

计 算 机 工 程
Computer Engineering

2011年11月
November 2011

·安全技术·

文章编号：1000—3428(2011)21—0102—03 文献标识码：A 中图分类号：TP301.6

基于特征相似度的贝叶斯网络入侵检测方法

王春东，陈英辉，常 青，邓全才，王怀彬

(天津理工大学计算机与通信工程学院，天津 300384)

摘 要：传统贝叶斯入侵检测方法未考虑属性和属性权值对检测结果的影响。为此，提出基于特征相似度的贝叶斯网络入侵检测方法。利用相似度对网络连接数据的属性特征进行选择，抽取其关键特征，并降低属性的冗余度，以优化朴素贝叶斯的分类性能。实验结果表明，该方法能降低分类数据的维数，提高分类的准确率。

关键词：特征选择；相似度；贝叶斯分类；入侵检测

Bayesian Network Intrusion Detection Method Based on Feature Similarity

WANG Chun-dong, CHEN Ying-hui, CHANG Qing, DENG Quan-cai, WANG Huai-bin

(School of Computer and Communication Engineering, Tianjin University of Technology, Tianjin 300384, China)

【Abstract】The traditional Bayesian intrusion detection method can not consider the fact that their different actions have differences between data attributes. This paper uses similarity to select the attribute features of network connecting data, gets the main feature, reduces attribute redundancy to improve the traditional Bayesian classification performance. Experimental results show that this method can reduce the dimension of the classification data, and improve the classification accuracy.

【Key words】feature selection; similarity; Bayesian classification; intrusion detection

DOI: 10.3969/j.issn.1000-3428.2011.21.035

1 概述

朴素贝叶斯网络入侵检测方法[1]是一种基于统计的检测方法，具有很强的概率推理判断能力，既可用来检测已知入侵，又可以检测未知或已知入侵的变种。但这种传统检测方法的缺点是在对未知样本进行分类时没有考虑不同属性对分类所起的作用不同[2-3]，即冗余的数据属性会提高数据的维度，增加分类计算量并带来噪音，造成分类准确性的下降。朴素贝叶斯中同一属性对于不同类别进行分类时相对作用的大小对分类准确度的影响也很大。文献[4]采用了一种建立在签名机制基础上的自动入侵监测系统，虽然能够有效地检测攻击，但是这一系统存在的主要困难就是难以选择参数。文献[5]采用贝叶斯方法来划分攻击的不同类型，可以有效地检测到攻击，并展现了这种方法的最大检测率和最小误报率，但是在有效性上还有待提高。

为了解决上述问题，本文在综合考虑网络入侵检测数据特点及贝叶斯分类方法优点的基础上，提出了基于特征相似度的贝叶斯入侵检测方法。

2 相似度的概念

对于不同的具体应用，其相似度的含义有所不同。例如，在基于实例的机器翻译中，相似度主要用于衡量文本中词语的可替换程度；而在信息检索中，相似度更多的是反映文本与用户查询在意义上的符合程度；而在多文档文摘系统中，相似度可以反映出局部主体信息的拟合程度。相似度的应用很广，在人工智能、生物医药、模式识别、数字图像、语义分析等方面都是重要的理论基础。

相似度是数学中的一个概念，用来判断2个数据样本之间的差异程度[6-7]，概率统计中的相关系数是一个很好的刻画2个随机变量分布相似程度的工具。n维变量X与Y的相关系数定义如下：

定义 1 设(X,Y)为二维随机变量，若$E\{[X-E(X)]\}\{[Y-E(Y)]\}$存在，则称它是随机变量$X$与$Y$的协方差，记为$Cov(X,Y)$，即：

$$Cov(X,Y)=E\{[X-E(X)][Y-E(Y)]\} \tag{1}$$

定义 2 若$D(X)>0$，$D(Y)>0$，称数值$\frac{Cov(X,Y)}{\sqrt{D(X)}\sqrt{D(Y)}}$为随机变量$X$与$Y$的线性相关系数，简称相关系数，记为$\rho_{XY}$：

$$\rho_{XY}=\frac{Cov(X,Y)}{\sqrt{D(X)}\sqrt{D(Y)}} \tag{2}$$

其中，$D(X)$，$D(Y)$分别为随机变量X和Y的方差。若$\rho_{XY}=0$，则称X与Y不相关，否则相关；若$\rho_{XY}>0$，称X与Y正相关；若$\rho_{XY}<0$，称X与Y负相关。ρ_{XY}越大，表明变量的相似度越高。

3 朴素贝叶斯分类

朴素贝叶斯分类的基本思想是[8]：通过计算后验概率的方法来确定样本所属类别的概率，将事件的先验概率与后验概率巧妙地联系起来，利用先验信息和样本数据信息确定事件的后验概率，最后把类归于后验概率最大的类别。

基金项目：天津市教委滨海双百基金资助项目(SB20080053，SB20080055)；教育部科技计划基金资助重点项目(208010)

作者简介：王春东(1969－)，男，教授、博士，主研方向：网络与信息安全；陈英辉、常 青、邓全才，硕士研究生；王怀彬，研究员

收稿日期：2011-04-20 **E-mail：**michael3769@163.com

根据贝叶斯公式，假设 $A_1, A_2, \cdots, A_n$ 是一组两两不相容的事件，而事件 B 能且只能与其中一个事件同时发生，那么则有下式成立：

$$P(A_i \mid B)=\frac{P(B \mid A_i)P(A_i)}{P(B)}=\frac{P(B \mid A_i)P(A_i)}{\sum_{i=1}^{n}P(B \mid A_i)P(A_i)} \tag{3}$$

朴素贝叶斯分类是基于一个简单假定，即对于给定的样本集，各个类别的概率分布是已知的，而且属性对类别的影响是相互独立的。换言之，该假定说明在给定实例的目标值情况下，观察到的联合概率就等于每个单独属性的概率乘积。即如果样本集有 n 个属性，即 $A_1, A_2, \cdots, A_n$ 属性值构成样本的特征向量，可能的类别有 m 个，具体为 $\{C_1, C_2, \cdots, C_m\}$。假设待分类样本 X 的特征向量为 $\{x_1, x_2, \cdots, x_n\}$，计算 X 分别属于每个类别的概率 $P(C_i \mid X)$，其最大值即为 X 的预测类别。其中，$P(C_i \mid X)$ 由下式计算得到：

$$P(C_i \mid X)=\frac{P(X \mid C_i)P(C_i)}{P(X)} \tag{4}$$

由于对于所有 C_i，$P(X)$ 都相同，因此只需要比较分子部分 $P(X \mid C_i)P(C_i)$ 即可。$P(C_i)$ 可由训练集得到，该值等于训练集类别为 C_i 的样本所占比例。在属性独立的假设下，则有下式成立：

$$P(X \mid C_i)=\prod_{j=1}^{n}P(x_j \mid C_i) \tag{5}$$

其中，$P(x_j \mid C_i)$ 由训练集估计得到。如果 A_j 是分类属性，$P(x_j \mid C_i)$ 等于类别为 C_i 的训练样本中属性 A_j，并且等于 x_j 的比例。则有下式成立：

$$P(C_i \mid X)=\frac{\prod_{j=1}^{n}P(x_j \mid C_i)P(C_i)}{P(X)}$$

又由于 $P(X)$ 都相等，因此只比较下式即可：

$$P(C_i \mid X)=P(C_i)\prod_{j=1}^{n}P(x_j \mid C_i) \tag{6}$$

在朴素贝叶斯网络的网络结构中，只有一个类结点，其他节点表示分类的各个属性，每个属性节点有且只有一个父节点，即类结点，且各个属性节点之间是相互独立的。

朴素贝叶斯分类结构如图 1 所示。

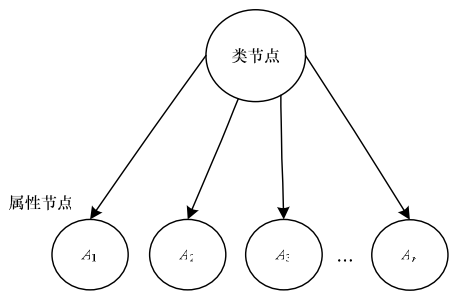

图 1 朴素贝叶斯分类结构

由图 1 可知，有向边代表节点间的概率依赖关系，朴素贝叶斯分类就是计算出各个节点的后验概率，以后验概率更新先验概率，最后取概率最大值作为分类的依据。

4 基于特征相似度的贝叶斯网络入侵检测方法

4.1 朴素贝叶斯分类方法改进

从理论上来讲，朴素贝叶斯分类比其他分类方法如决策树等具有更好的分类精度，但是贝叶斯分类模型对样本进行分类时没有考虑属性不同对分类所起作用也不同，冗余的数据属性会提高数据维数，会增加分类计算量并且带来噪声影响，从而造成分类准确率的下降。对于这种情况，可以在分类之前对样本进行特征选择。常用的特征选择方法有信息增益、卡方检验、统计方法等，虽然采用这些特征选择方法可以提高分类性能，但是其只考虑了特征属性 A_i 与各个类别 $C_i(1 \leqslant i \leqslant m)$ 之间相关度之和的最大值，而这个最大值并不能全面衡量各个特征属性 A 对分类意义的影响大小。因此，应该考虑各个属性之间的相关性，和各个属性特征间彼此联系的缜密性如何。

鉴于此，本文引入相似度这一概念，提出基于相似度的特征选择方法，通过此方法抽取关键特征来降低属性的冗余度，达到降低数据难度的目的，这样既可以保证正确获得数据的特征，又能减少计算量和提高分类效率。

相似度的具体方法如上述式(1)、式(2)所示，即把式(2)运用到朴素贝叶斯方法中，以得到本文方法：

$$P(C_i \mid X)=\arg\max P(C_i)\prod_{j=1}^{n}\rho_j P(x_j \mid C_i) \tag{7}$$

4.2 本文方法的实现及分析

本文实验采用的训练集和测试集均来自入侵检测领域比较权威的 KDD CUP 1999 Data Sets[9]，此数据集包含 490 万条数据记录。所有的攻击主要分为 4 大类：DOS 类，PROBE 类，R2L 类和 U2R 类。其中，DOS 类中包含的入侵类型有 Land、Neptune、Pod、Teardrop 等；PROBE 类包含的入侵类型有 Nmap、Portsweep、Satan、Mscan、Ipsweep 等。

为了处理方便，实验采用整个数据集的 10%，即大约 10 万条记录作为样本，取其中的 7 万条作为训练集，其余的 3 万条作为测试集，实验步骤如下：

(1)对于训练样本集，首先进行预处理操作。由于 NB 方法要求所有的数据必须是离散值或者是经过离散化的值[10]，因此实验的首要任务就是把所有的数据离散化。根据这些数据的特点，本次实验采用数据挖掘工具 WEKA 自带的离散化工具进行离散化。

(2)利用式(2)计算属性 A_i 相互之间的相似度，选择低于阈值的属性值，根据实验经验，阈值选择 0.95。删除高于阈值的属性后，得到一个新的属性集合 $B=\{B_1, B_2, \cdots, B_k\}$，$k \leqslant n$，$n$ 是训练集中属性总个数。通过这个过程定性、定量的衡量属性间的相似程度，删除冗余属性，达到降维目的。

对于数据源中的 Smurf 攻击，计算 ρ(dst_host_srv_rerror_rate,dst_host_rerror_rate)=100%，1>0.95，故超过阈值，即两者中有其一为冗余属性。数学中方差反映了变量与其均值的偏离程度，其偏离越大，则发生特殊情况的可能性就越大，认为其有冗余属性，故选择方差较小的属性。本例中的 2 个属性，方差相同，则选择其一即可。经过特征选择得出 Smurf 攻击用于分类的特征属性如下：protocol，service，flag，src_bytes，count，srv_count，srv_rerror，dst_host_count，dst_host_srv_count，dst_host_same_src_port_rate，dst_host_ srv_rerror_rate。其他类型的攻击的属性选择方法与此类似。

(3)计算 $P_i=P(C_i)$。表示为类 C_i 的样本在集合中出现的概率。

(4)对于步骤(2)中得到的最优属性集合 B_k，计算每个属性 B_k 的条件概率 $P(X \mid C_i)=\prod_{j=1}^{k}P(x_j \mid C_i)$，其中，$P(x_j \mid C_i)$ 等于类别为 C_i 的训练样本中，属性 B_j 等于 x_j 的概率。

(5)利用下式进行分类：

$$P(C_i \mid X) = \arg\max P(C_i)\prod_{j=1}^{n}\rho_j P(x_j \mid C_i)$$

5 实验结果与分析

为了突出本文方法的优越性，将其与传统的贝叶斯方法和神经网络方法进行比较。调用 WEKA 分类器(版本采用的 3.5.7)当中的 Bayesian 方法来获得数据集在传统的贝叶斯方法的准确率。评估指标：检测率=检测到的攻击/总的攻击数量。检测率对比结果如表 1 所示。

表 1 检测率比较 (%)

攻击名称	传统贝叶斯方法	神经网络方法	本文方法
Normal	92.95	90.11	96.25
Neptune	89.10	86.40	92.70
Satan	80.70	79.40	82.70
Mscan	80.30	78.80	93.60
Land	95.30	95.20	97.50

由表 1 可知，在引入相似度以后，与传统的贝叶斯方法和神经网络方法比较，检测率有明显的提高。与前两者比较，本文方法是一种基于概率统计的方法，具有良好的概率推理机制，即改进的方法更加科学可靠。

6 结束语

本文提出基于特征相似度的贝叶斯网络入侵检测方法。综合考虑网络入侵检测数据特点及贝叶斯分类方法的优点，删除冗余属性特征，将概率统计中的相关系数引入到本文方法中。

实验结果证明，该方法与传统贝叶斯方法和神经网络方法相比，检测效率有所提高，并且降低了分类数据的维数。

参考文献

[1] 薛静锋. 基于数据挖掘入侵检测的研究[D]. 北京: 北京理工大学, 2003.

[2] 王 树, 杜启军, 余桂贤. 网络入侵检测系统的最优特征选择方法[J]. 计算机工程, 2010, 36(15): 140-141, 144.

[3] 郑洪英, 侯梅菊, 王 渝. 入侵检测中的快速特征选择方法[J]. 计算机工程, 2010, 36(6): 262-264.

[4] Farah J, Montaceur Z, Mohamed B A. Intrusion Detection Based on “Hybrid” Propagation in Bayesion Networks[C]//Proc. of IEEE Intelligence and Security Informatics Conference. [S. l.]: IEEE Computer Press, 2009.

[5] Farid D M, Rahman M Z. Learning Intrusion Detection Based on Adaptive Bayesian Algorithm[C]//Proc. of Conference on Computer and Information Technology. [S. l.]: IEEE Computer Press, 2008.

[6] 梁冯珍, 宋占杰, 张玉环. 应用概率统计[M]. 天津: 天津大学出版社, 2005.

[7] 唐振江, 何 慧, 云晓春. 基于多特征相似度的蠕虫检测[J]. 高技术通讯, 2005, 15(8): 11-17.

[8] 李柏生, 林亚平. 基于卡方检验的贝叶斯网络入侵检测的分析[J]. 计算机工程与设计, 2008, 29(15): 3849-3851.

[9] Stolfo S, Wenke L. KDD Cup 1999 Data[EB/OL]. (2010-11-21). http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html.

[10] Wojciech T. Anomaly-based Intrusion Detection Using Bayesian Networks[C]//Proc. of the 3rd International Conference on Dependability of Computer Systems. [S. l.]: IEEE Computer Press, 2008.

编辑 刘 冰

(上接第 101 页)

δ-HSSVM 表示使用一个 δ-HSSVM 对数据进行检测。因为 δ-HSSVM 只处理同维数的数据，所以对 UDP 和 ICMP 数据补 0，使 UDP、ICMP 数据与 TCP 数据的维数相等。Multiδ-HSSVM 表示用本文的协同方法对数据进行检测。由表 1 可以得到 Multi δ-HSSVM 方法总的结果：*TN*=266+189+298=753；*TR*=753/900=83.67%；*R-error*= 2.61%；*T*−*time*≤1.82 s。

表 2 2 种检测方法的检测结果对比

检测方法	*TN*	*TR*/(%)	*R-error*/(%)	*T*−*time*/s
Single δ-HSSVM	691	76.78	5.57	7.60
Multi δ-HSSVM	753	83.67	2.61	≤1.82

从表 2 的数据可以看出，本文协同方法 Multi δ-HSSVM 比 Single δ-HSSVM 方法有效，特别是在检测时间 Multi δ-HSSVM 要远低于 Single δ-HSSVM 时。

从上面的分析可以看出，因为 UDP 和 ICMP 数据集是不平衡数据集，致使包裹这些数据集的超球把较多的数据点视为噪声放在球的外部，从而 Single δ-HSSVM 的 *TR* 减小，*R-error* 增大。且在测试阶段有大量的样本处在 2 个超球重叠处，故 *T-time* 增大。

6 结束语

本文的研究 HSMCSVM 算法，并对其进行改进，提出基于 δ-HSSVM 的协同入侵检测方法。实验结果表明，该入侵检测方法能取得较好的检测效果。进一步的研究工作是如何增大 δ-HSSVM 的分类准确率以及构建适当的核函数，使超球重叠数减少。

参考文献

[1] 徐 蕾, 刘冬好. 基于层次决策表增量学习算法的网络入侵检测[J]. 计算机工程, 2010, 36(17): 173-175.

[2] 朱美琳, 刘向东, 陈世福. 用球结构解决多分类问题[J]. 南京大学学报, 2003, 39(2): 153-158.

[3] 吴 强, 贾传荧, 张爱峰, 等. 球结构支持向量机的改进算法及仿真研究[J]. 系统仿真学报, 2008, 20(2): 345-348.

[4] Teng Shaohua, Du Hongle, Wu Naiqi, et al. A Cooperative Network Intrusion Detection Based on Fuzzy SVMs[J]. Journal of Networks, 2010, 5(4): 475-483.

编辑 刘 冰