

一种基于代价敏感学习的范例推理方法及其应用研究

罗菲菲,刘贵全,安景琦,张婷慧

(中国科学技术大学 计算机科学技术系,安徽 合肥 230027)

(ffluo@mail.ustc.edu.cn)

摘 要:提出一种基于代价敏感学习的范例推理方法,可以对大规模、高维数据进行分类和预测。该算法在分类的同时,不断调整数据属性项权重,以减少由分类引起的误分代价。在某入侵检测数据分析中取得了较好的结果。

关键词:范例推理;分类算法;代价敏感

中图分类号: TP18 **文献标识码:** A

Research on a case-based reasoning method using cost-sensitive learner and its applications

LUO Fei-fei, LIU Gui-quan, AN Jing-qi, ZHANG Ting-hui

(Department of Computer Science, University of Science & Technology of China, Hefei Anhui 230027, China)

Abstract: This paper proposed a case-based reasoning method using cost-sensitive learning that can classify and forecast large-scale and high dimension data. While classifying, this method adjust attribute weight constantly, in order to reduce the misclassification cost. The method has made a better result on some intrusion detection research.

Key words: case-based reasoning; classify algorithm; cost-sensitive learning

0 引言

基于范例推理(CBR)是人工智能的一项技术,它利用已有的知识(即源范例)来解决当前遇到的新的问题(即目标范例)。与基于规则的推理方法相比,范例的获取比规则获取容易,可以降低知识获取的难度。范例推理方法能够对过去的求解结果进行复用,提高对新问题的求解效率。现有的入侵检测系统(IDS),通常采用基于规则的入侵检测方法,将入侵行为用规则表示,通过将系统行为与入侵规则的匹配来发现攻击行为。这种一对一的规则匹配方法缺乏一定的灵活性,对于检测未知的攻击行为存在很大的困难。基于范例推理的入侵检测方法,采用近似推理的思想,在知识获取、知识更新、系统的自学习和扩展能力上都优于基于规则的入侵检测方法。

范例由范例的特征项来描述,不同特征项对范例的重要性不同,通常使用特征项权重来描述范例中各个特征项的重要性。特征项权重在范例推理中非常重要,它涉及范例间相似性计算以及范例索引等。范例特征项权重学习对范例的检索质量和检索速度起着非常重要的作用。在IDS中,范例间相似性计算及范例查询将导出范例的分类问题。

在代价敏感学习方面已经有了很多的研究。Chan 和 Stolfo 提出了 meta-learning 方法可以降低由于信用卡欺诈检测所引起的系统代价^[3]。Breiman 提出 MetaCost 方法,利用 bagging(自助聚类)算法,来估计分类概率^[4]。在特征项代价敏感研究中,Lavrac 等人采用遗传算法有效的解决特征项过滤问题^[2]。Wenke Lee 等人对几种入侵行为进行了详细分

析,并在此基础上给出了相对应的损失代价,响应代价和操作代价^[1]。

与其他的研究成果相比,本文提出的基于代价敏感学习的范例推理方法,是一种针对范例特征项值对权重调整的具体算法。算法在进行权重调整时,采用了 Wenke Lee 等人对入侵行为损失代价分析的结果,并且将其分析结果应用到具体的实验中。实验表明,基于代价敏感学习的范例推理方法,能够让系统自适应性的调整范例的特征项值对权重,达到降低系统误分代价的目的。目前采用代价敏感学习来调整范例特征项权重的研究还比较少。基于代价敏感学习的范例推理方法,不仅适用于入侵检测系统,同时也适用于其他需要考虑误分代价的系统

分类问题是数据挖掘和机器学习领域内的一个重要问题,主要的分类算法有决策数、神经网络、贝叶斯算法等。这些算法都要追求尽可能小的误分率,即对样本的误分次数。然而在现实情况下,不同类之间的误分所引起的代价并不完全相同。例如,IDS 将 U2R 攻击误分为 Probing 和误分为 Normal 所引起的决策代价显然是不同的,因而单纯考虑误分率,而不考虑系统的误分代价是不合理的。本文采用基于代价敏感的学习的范例推理方法,该算法在得到较低的查询范例误分率时,能够尽量减少系统的误分代价。

1 背景与研究思路

1.1 范例推理方法

范例由范例的特征项及特征项值来描述。假定用 n 个特征项 $F_i (i = 1, 2, \dots, n)$ 来描述范例,则对于范例 x , 范例 x 表

收稿日期:2005-04-19;修订日期:2005-07-11

作者简介:罗菲菲(1980-),女,四川成都人,硕士研究生,主要研究方向:范例推理、入侵检测、数据挖掘; 刘贵全(1970-),男,四川人,副教授,博士,主要研究:自然语言理解、网络安全; 安景琦(1981-),男,安徽人,硕士研究生,主要研究方向:入侵检测; 张婷慧(1982-),女,湖北人,硕士研究生,主要研究方向:数据挖掘。

达为:

$$x = \{F_1:x_1, F_2:x_2, \dots, F_n:x_n, c_x\}。$$

其中 $x_i (i = 1, 2, \dots, n)$ 分别为 n 个特征项对应的值, x 所属的类别为 c_x 。

在范例推理中,范例间距离的计算是非常重要的,它是范例间相似度的评价和范例检索等的基础。距离的计算方法有很多种,本文采用 Minkowski 距离来计算:设查询范例 q 与范例 x 之间的距离用 $d(q, x)$ 表示,则计算公式为:

$$d(q, x) = \left(\sum_{f \in F} w(F) \cdot \delta(x_F, q_F)^r \right)^{\frac{1}{r}}$$

其中, $w(F)$ 为特征项权重函数, $\delta(\cdot)$ 是依不同特征项类型定义的值函数。当特征项为连续值时, $\delta(x_F, q_F) = |x_F - q_F|$; 当特征项为字符时, $\delta(x_F, q_F) = 0$ if $x_F = q_F$, $\delta(x_F, q_F) = 1$, if $x_F \neq q_F$ 。

范例检索有很多方法,本文采用 k -近邻法作为范例检索依据:令 $P(c_i | q)$ 表示将范例 q 分类为 c_i 的概率。对于查询范例 q 所属类别,采用最大可能所属类别概率 $p(c_j | q)$ 表示。假设与 q 最近的 k 个近邻的类别集合为 C_k , 则范例 q 所属类别 c_q 定义为:

$$c_q = \arg \max_{c_j \in C_k} p(c_j | q)$$

1.2 代价敏感学习方法

分类算法有时很难得到完全正确的分类结果,错误的分类将引入误分代价问题。采用代价敏感学习方法,可以有效减小系统的误分代价。

对于范例 q , 假定 q 的类别为 c_j , 令 $P(c_i | q)$ 为将范例 q 分类为 c_i 的概率, L_{c_j, c_i} 表示将 c_j 类的范例误分为 c_i 类的代价。定义 $R(c_i | q)$ 为将 q 分为类 c_i 的代价, 则 $R(c_i | q) = L_{c_j, c_i} \times P(c_i | q)$ 。 $R(q)$ 表示由 q 的分类问题引起的系统误差, 定义 $R(q) = \sum_{c_i \in C} R(c_i | q)$ 。此时系统总的代价为 $E(w) = \sum_{q \in Q} R(q)$, 即 $E(w) = \sum_{q \in Q} \sum_{c_i \in C} L_{c_j, c_i} \times P(c_i | q)$ 。

对于一般的对称损失而言, 当 $c_j = c_j$ 时 $L_{c_j, c_j} = 1$, 当 $c_j \neq c_j$ 时 $L_{c_j, c_j} = 0$; 但是在入侵检测中, 将一个对象误分为不同类所引起的误分代价是不同的, 是一个代价敏感的问题。在入侵检测系统中有诸多代价因素, 例如损失代价, 响应代价, 操作代价等, 由于损失代价是最主要的部分, 本文主要考虑由误分引起的损失代价。

2 基于代价敏感的范例特征项学习算法

2.1 算法描述

范例的特征项与其特征项取值存在着——对应的关系, 本文采用特征项对这个概念来描述范例的特征项与其对应的特征项取值。假设范例库 CP 中有 M 个范例, 每个范例有 N 个特征项, 则相对于范例库 CP 而言, 每个特征项 $F_i (i = 0, 1, \dots, N)$ 的取值范围是一定的。假设特征项 F_i 有 $m_i (i = 0, 1, \dots, N)$ 个不同的取值, 则特征值对的总数为 $n = \sum_{i=1}^N m_i$, 此时标记各个特征值对为 $F_i x_j (1 \leq j \leq n)$ 。如果范例 p 的特征 F_i 的取值为 x_j , 则称 p 与特征值对 $F_i x_j$ 关联。当范例 p 与特征值对 $F_i x_j$ 之间有关联时, 则用特征值对权重 w_{pij} 表示其关联强度。

特征值对的权重反映了不同特征值对相对于范例的重要

性。如果某些特征值对经常被用来表征某个范例, 则它们之间的权重相对较大。这一点与神经网络中的权值变化非常类似, 当一个神经元不断被刺激时, 其联接权重就不断增大, 反之, 权重衰减。基于代价敏感的范例特征项学习算法利用代价敏感矩阵来调整 w_{pij} , 使得调整后的特征项权重能准确的反应它对范例的重要性程度, 使查询范例的系统误分代价尽可能小。

2.2 算法实现

算法 1 范例库训练算法

输入: 源范例库 CP, 训练范例库 CQ, 初始权重 W , 学习速率 η , 代价敏感矩阵 M 。初始计数 $count = 0$ 。

1) 数据预处理: 范例特征项过滤。由于范例的特征项较多, 且存在一些无关和冗余的特征项, 对范例的特征项过滤, 可以提高算法的运行效率。

2) 针对 CQ 里的每一个训练范例 q , 用 k -近邻法作为范例检索依据, 计算它与每一个范例库 CP 中的范例 p 的距离 $d(q, p)$, 以及最大可能所属类别概率 $p(c_j | q)$, 得到范例 q 所属类别 c'_q 。 c'_q 为范例 q 的预测类别。定义此时:

$$d(q, p) = \left(\sum_{f \in F} w_{pij} \cdot \delta(p_{Fi}, q_{Fi})^r / \sum_{f \in F} w_{pij} \right)^{\frac{1}{r}}$$

3) 比较 c'_q 与范例 q 自身的类别 c_q , 调整范例 q 的 k -近邻的范例 p 的特征项权重 w_{pij} , 调整后的特征项权重用 w'_{pij} 表示。

$$w'_{pij} = \begin{cases} w_{pij} (1 - G(M_{c'_q, c_q}) \times t \times f), & \text{if } c'_q \neq c_q \\ w_{pij} \times (1 + \eta \times f), & \text{if } c'_q = c_q \end{cases}$$

其中当特征项取值为符号值时:

$$f = \begin{cases} 1, & \text{if } p_{Fi} = q_{Fi} \\ 0, & \text{if } p_{Fi} \neq q_{Fi} \end{cases}$$

当特征项取值为连续值时:

$$f = \frac{1}{1 + |p_{Fi} - q_{Fi}|}$$

$G(M_{c'_q, c_q})$ 为代价敏感函数, η 为学习速率, t 为调整因子, p_{Fi} , q_{Fi} 分别为范例 p, q 在第 i 个特征项 F_i 上的取值。

4) 重复 2), 3) 直到将所有的训练数据分类。然后分别计算 w_{pij} 调整前后的系统查询误差 $E(w)$ 和 $E(w')$ 。

5) 如果 $|E(w) - E(w')| < \varepsilon'$, 或者 $E(w') < \varepsilon$ 或者 $count \geq N$ 则转到 6); 否则重复 2) ~ 4), $count++$ 。

6) 其中 N 为人为设定的最大可接受循环次数, ε 为最大可接受查询误差。若此时 $E'(w) < \varepsilon$, 则算法结束, 否则调整学习速率 η , 重复步骤 1) ~ 5);

输出: 每次迭代后的查询范例识别率和查询误差 $E(w)$ 。

在算法中, 当用训练范例进行范例检索时, 得到预测类别。如果预测类别正确, 则增加 k -近邻的范例 p 中相关特征值对权重 w_{pij} ; 如果预测类别错误, 则减少 p 的相关特征项权重, 减少的幅度由误分代价决定。误分代价越高, 幅度越大。这样可让算法尽快收敛到正确的结果。

算法 2 范例库测试算法

输入: 源范例库 CP, 测试范例库 CR, 调整后的权重 W 。初始计数 $count = 0$ 。

1) 为 CR 里的每一个测试范例 r , 用 k -近邻法得到 k 近邻, 计算最大可能所属类别概率 $p(c_j | r)$, 得到范例 r 的预测类别 c'_r 。比较 c'_r 与 c_r , 如果 $c'_r = c_r$, 则 $count++$ 。

2) 计算测试范例的识别率, 识别率 = $\text{count} / N(CR)$, $N(CR)$ 为范例库 CR 的范例数。

3 计算此时的系统误差 $E(W)$ 。

输出: 测试范例的识别率及系统误差 $E(W)$ 。

3 实验结果与分析

3.1 实验数据

取自 KDDCUP99 提供的实验数据, 随机抽取一部分, 共计 2 422 条数据, 23 个攻击类别。每条数据有 42 个属性项, 其中字符属性 8 个, 连续值属性 34 个。将数据按 DARPA 给出的分类方法将 23 类攻击映射为 5 大类 (U2R, R2L, DoS, Probing, Normal)。实验数据分成 3 部分, 源范例库, 训练范例库和查询范例库。其中, 源范例库数据占 50%, 训练范例库 30%, 查询范例库数据占 20%。

3.2 误分损失代价矩阵

文献[1]给出了四种不同攻击类型的损失代价取值, 将损失代价扩展得到误分代价矩阵。扩展方法为, 对于范例类型 A 和 B , 定义 $\text{Cost}(A)$ 为 A 的损失代价, $\text{Cost}(A, B)$ 为将 A 误分为 B 的代价; 其中 $\text{Cost}(A, B) = \text{Cost}(B, A)$; if $\text{Cost}(A) > \text{Cost}(B)$, 则 $\text{Cost}(A, \text{normal}) = \text{Cost}(A, B) + \text{Cost}(B, \text{normal})$ 。

攻击类别	损失代价
U2R	100
R2L	50
DoS	30
Probing	2
normal	0

误分损失代价矩阵

误分代价	查询范例的预测类别				
正确类别	Normal	DoS	U2R	R2L	Probing
Normal	0	0	0	0	0
DoS	30	0	70	20	28
U2R	100	70	0	50	98
R2L	50	20	50	0	48
Probing	2	28	98	48	0

图 1

3.3 实验结果

1) 训练结果

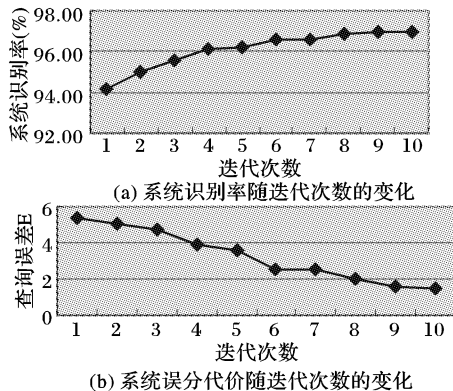


图 2 实验结果

初始时, 查询范例的识别率为 94.11%。 $E(w) = 5.361224$ 。取 $\eta = 0.2$, $t = 0.008$ 经过 10 次迭代之后, 查询识别率达到 96.99%, 查询误差 E 降低到 1.47。具体结果如图 2

所示。

从实验结果可以看出, 基于敏感代价的范例特征项权重调整算法, 在实验中呈现出很好的规律性, 即随着迭代次数增加, 识别率逐渐提高, 而误差逐渐降低。在迭代 9 次之后识别率稳定在 97% 以上, 查询误差稳定在 1.5 左右。

2) 测试结果

测试范例的识别率为 96.57%, 查询误差 $E = 1.73$ 。测试结果表明, 经过实验 1 训练得到的特征值对权重能够有效的反映各特征值对的重要性, 为范例检索提供了可靠的依据, 同时系统的查询误差也较低。

4 结语

基于规则匹配的入侵检测方法存在着一定的局限性, 如果规则库里的规则不够充分, 则有可能无法准确的检测出入侵行为。基于范例推理方法采用近似推理的思想, 在知识学习及发现上都优于基于规则的入侵检测。与 ID3 等判定树分类算法比较, 本文提出的方法, 可以适用于同时含字符属性和连续值属性的数据, 不需要对数据进行离散化, 在数据处理上更为精确。同时引入代价敏感矩阵, 针对特定类型的数据, 该算法能够更为有效的降低误分代价。在实验中, 算法经过有限次迭代后, 训练范例的识别率稳定在 97% 以上, 查询误差可以降到在 1.5 以下, 测试范例的识别率为 96.57%, 查询误差为 1.73。实验结果表明, 基于代价敏感学习的范例推理方法, 可以有效的提高系统的识别率, 同时降低系统的误分代价。

在算法描述和实验部分, 我们可以看到, 算法的关键在于权重调整的方法和代价矩阵的设计。这两点对于实验最终的结果影响非常的大。进一步的研究方向包括:

1) 除了考虑系统的损失代价外, 引入系统的响应代价和操作代价, 对入侵行为对系统总的影响进行考虑, 使算法能够更为全面真实的反映系统总的代价。此后根据实际情况调整代价矩阵, 使之更加合理和有效。

2) 研究更为有效的特征项调整方法, 提高算法的效果和效率。

参考文献:

- [1] LEE W, WEI F, MILLER M. Toward Cost-Sensitive Modeling for Intrusion Detection and Response[A]. Workshop on Intrusion Detection and Prevention, 7th ACM Conference on Computer Security [C]. Athens, GR: November, 2000.
- [2] LAVRAC N, GAMBERGER D, TURNEY P. Cost-Sensitive Feature Reduction Applied to a Hybrid Genetic Algorithm[A]. In Proceedings of the Seventh International Workshop on Algorithmic Learning Theory[C], Sydney, Australia, 1996. 127 - 134.
- [3] CHAN P, STOLFO S. Towards scalable learning with non-uniform class and cost distribution: A case study in credit card fraud detection[A]. In Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD-98) [C], August, 1999.
- [4] DOMINGOS P. MetaCost: A general method for making classifiers cost-sensitive[A]. In Proc. of the Fifth ACM SIFKDD int'l. Conf. on Knowledge Discovery Data Mining[C]. San Diego, CA, ACM, 1999. 155 - 164.