基于快速属性选择的贝叶斯分类在入侵检测中的应用*)

王 翔 胡学钢
(合肥工业大学计算机与信息学院 合肥 230009)

摘 要 高速网络环境中数据量日益增大,安全问题日益突出,对入侵检测技术提出了更高的要求。朴素贝叶斯作为数据挖掘的重要方法之一,在入侵检测中有着重要的地位。由于其属性独立假设,使得如何在海量高维数据处理背景下快速、准确、有效地选出代表原数据的属性显得尤为重要。本文提出了一种快速属性选择方法并结合朴素贝叶斯分类模型应用于入侵检测中。实验表明,结合了该属性选择方法的朴素贝叶斯分类器有很好的分类精度及较低的时空消耗。

关键词 快速属性选择,朴素贝叶斯分类,入侵检测

Bayesian Classifier Based on the Fast Attribute Selection for Intrusion Detection

WANG Xiang HU Xue-Gang
(School of Computer and Information, Hefei University of Technology, Hefei 230009)

Abstract As a consequence of the increasing security problems and huge data in high-speed networks, it is urgent to develop new Intrusion Detection Techniques. Naïve Bayesian is one of the most important approaches in data mining, and it has a high ranker in Intrusion Detection. According to its conditional independence assumption, it is very critical to select attributes, which represent raw data, quickly, and effectively when processing huge high-dimension data in networks. This paper proposes a Fast Attribute Selection idea combined with Naïve Bayesian Classifier, which is used in Intrusion Detecting. Experiment's results show that our idea has high precision and low cost.

Keywords Fast attribute selection, Naïve Bayesian classifier, IDS

高速网络的普及使得人们充分享受了网络带来的便捷,随着网络安全问题的日益突出,对网络的各类攻击与破坏已经成为学术界、企业、政府部门共同关心的事件。为保护计算机网络的安全,入侵检测系统等的建立是十分必要的。

当前用于入侵检测的数据挖掘算法有神经网络、决策树、序列模式、关联规则等比较经典的算法。贝叶斯分类方法具有空间复杂度低,时间复杂性仅为线性等优点,在海量数据处理方面已表现出高准确率与高速度[4]。朴素贝叶斯分类器是基于贝叶斯分类方法的分类器,是一种简单的、有效的、在入侵检测中有着成功应用的重要分类器。它假定某一个属性值对给定类的影响独立于其他属性的值,简化了计算。实际应用中,如何从高维海量网络数据中快速、有效地选择满足其独立假设的属性成为重要的研究内容[1]。目前针对入侵检测数据预处理获得约简属性有很多方法,如基于粗糙集合的属性约简[5,7,9],Relief 方法[6],Wrapper 方法等等。入侵检测数据海量高维的特点使得这些需要大量计算的属性选择方法直接应用时面临时空开销巨大等难题。

本文在综合考虑入侵检测数据特点及朴素贝叶斯分类器优点的基础上提出了一种快速属性选择方法,利用统计方法,尽量减少复杂计算,从而实现快速选择代表当前数据特征并能提升朴素贝叶斯分类器性能的属性。

1 朴素贝叶斯

1.1 朴素贝叶斯分类过程

贝叶斯分类的过程如下[1]:

(1)每个入侵检测数据样本用一个 n 维特征向量 $x=\{x_1,x_2,\cdots,x_n\}$ 表示,分别描述对 n 个属性 $A_1,A_2,\cdots,A_n$ 的度量。

(2)假定有 m 个类 $C_1,C_2,\cdots,C_m$。给定一个未知的数据样本 x,分类法将预测 x 属于具有最高后验概率的类,满足条件

$$\exists i, p(c_i|x)=\max_j(p(c_j|x)), 1\leqslant i,j\leqslant m\ i\neq j$$

则 x 属于类别 C_i。根据贝叶斯定理:

$$\max_j(p(c_j|x))=\max_j p(x|c_j)p(c_j)/p(x) \quad (1)$$

(3) 类的先验概率可以用 $p(c_j)=S_j/S$ 计算其中 S_j 是类 C_j 中的训练样本数,而 S 是训练样本总数。$p(x)$是常数,与决策分类无关,关键是如何计算 $p(x|c_j)$。由于朴素贝叶斯在给定类下,属性独立的假设,即在属性间,不存在依赖关系,从而在训练数据集中计算 $p(x_1,x_2,\cdots,x_n|c_j)$的值简化为:

$$p(x_1,x_2,\cdots,x_n|c_j)=p(x_1|c_j)p(x_2|c_j),\cdots,p(x_n|c_j) \quad (2)$$

(4) 对于未知样本 x 分类,对每个 C_j 类,计算 $p(x|c_j)$。样本 x 被划分为类 C_j 当且仅当:

$$p(x|c_j)p(c_j)>p(x|c_i)p(c_i), 1\leqslant i\leqslant m, i\neq j$$

1.2 分析

从理论上分析,朴素贝叶斯具有比其他分类算法如决策树、SVM 等更好的分类精度,但是由于现实数据很难保证其属性独立假设的前提,因而限制了它的发挥。尽管如此,大量实验表明朴素贝叶斯分类器依然具有高分类精度和较好的健壮性[3]。如选择代表原数据特征并满足朴素贝叶斯属性独立假设的属性进行训练,其分类精度可进一步提升。

2 快速属性选择算法

2.1 离散化

*)本文受安徽省自然科学基金课题(编号 050420207)资助。王 翔 硕士研究生,研究方向:数据挖掘,人工智能;胡学钢 博士,研究方向:知识工程,数据挖掘,人工智能。

在进行属性选择前，需要对原数据做预处理，将连续属性离散化。在考虑入侵检测数据特点和快速属性选择的要求后，采用如下方法进行离散化处理：

(1)对于其中的“开关”属性，即取值仅为0或1的连续属性，把它作为离散属性处理。

(2)对于其中记录的数据流量及变化率的连续属性，按照划分等宽区间的方法进行离散化，划分区间的标准为类别属性取值的个数。如：类属性 C 有 m 个取值，连续属性 A_i 的离散化方法为，计算 $\max(A_i)$ 和 $\min(A_i)$，区间宽度为 $(\max(A_i) - \min(A_i))/m$。

2.2 快速属性选择算法

理想情况下，若某属性 A_i 有 m 个取值 $v_1, v_2, \cdots, v_m$，类别属性 C 也有 m 个取值 $C_1, C_2, \cdots, C_m$，且 A_i 的取值与 C 的取值为对应关系，即属性 A_i 中的一个取值 V_i 可单独决定一个类 C_j，$1 \leqslant i, j \leqslant m$，则此属性 A 为最佳分类属性。由此推出假设，越接近上述这样的属性，其分类效果越好，与其他属性间独立性也越强。

基于上述思想，本文提出了如下的快速属性选择算法：

输入：数据样本集 T，其中单个样本 $x=\{x_1, x_2, \cdots, x_n\}$ 表示 $A_1, A_2, \cdots, A_n$ 的度量
输出：选择出的属性 Attrs{}
对训练数据集合 T 数据预处理(离散化)；
Attrs={ }；initial=20%实例数
while A≠Ø do
if 属性集合 $A=(A_1, A_2, \cdots, A_n)$ 中属性 $A_i=(v_1, v_2, \cdots, v_m)$ then 依次扫描 A 取值
for i =1 to A_i 取值个数
if $A_i=v_i$ then 扫描对应的类属性取值
count($C=C_i$)；//计算 $A=v_i$ 下各个类的取值个数
errors=C_i 总取值个数 − max(count($C=C_i$))；//计算误分类个数
if (initial- errors)/errors <λ then //停止搜索条件
Attrs=Attrs∪{A_i}
$A=A-\{A_i\}$；
initial = errors；
end
if 安全专家建议的属性 A_j not in Attrs then do //为专家经验提供的重要属性放松阈值条件，避免重要属性漏选。
if (errors(A_j) −worst(Attrs))/worst(Attrs) < ∂ then Attrs=Attrs∪{A_j}
end

2.3 算法分析

(1)初始化及阈值 λ、∂ 设定

初始化的预计错误率及阈值 λ 的设置对于选择的属性数量有很大影响。初始化预计错误率设为20%及算法中搜索停止阈值 λ 设定为10%，这两个值均为实验获得。阈值∂对于领域知识提供的属性能否入选最终的属性集很关键，经过大量实验得到阈值∂为20%时最多的专家属性入选且使得所选择出的属性总个数较少，速度较快且能保证朴素贝叶斯分类器的分类精度。

(2)算法的调整

该算法选择出的属性虽然接近属性独立且符合原数据特征，但为防止出现过拟合情况，结合网络安全专家给出的入侵检测数据中较重要的属性如 service(网络服务)，flag(连接状态)，src_bytes(源主机到目的主机的字节数)，dst_bytes(目的主机到源主机的字节数)，dst_host_srv_count(目标主机服务的数量)，diff_srv_rate(不同服务连接占的百分比)[7]，确保了更符合原入侵检测数据特征且提高了贝叶斯分类器的精度。

(3)性能分析

设原数据有 N 个属性，共有 m 条数据记录，则属性选择算法时间复杂性与样本数成线性关系 $O(mN)$。由于选择属性时需要多遍扫描数据库，初始化时设计为将数据文件读入内存进行处理，因而空间消耗与样本的大小接近正比。使用朴素贝叶斯分类时，时间消耗与样本数量 e 及训练样本属性 n 成正比为 $O(en)$，实验设计每次处理2000条数据，因此内存消耗仅与这2000条记录及分类模型大小有关，详见实验部分。

3 实验与分析

实验采用 KDDCUP99 数据集[8] 作为实验数据来源，该入侵检测数据由 MIT 林肯实验室提供，来源于真实网络数据，共有42个属性，其中34个连续属性，7个离散属性，1个类别属性。KDDCUP99-10percent 训练数据来源，包含 494,020 条记录；属性选择方法对比性实验中由于 WEKA 软件内存消耗限制，训练数据使用的是从 KDDCUP99-10percent 中随机抽取的 50,000 条记录；测试数据为 KDDCUP99-10percent 中随机抽取的 200,000 条记录。测试环境是基于赛扬 2.66G 内存 512M 的 PC 机，算法实现环境为 Windows XP Professional，Visual C++6.0。(注：WEKA 是新西兰 Waikato 大学开发的开源数据挖掘系统，包含很多经典分类算法，如 J4.8 即 C4.5 等等。)

实验在比较 WEKA 的常用属性选择方法基础上，Infogain，Wrappersubset+J48，Gainratio，ReliefAttribute 等，根据其时间和空间消耗大小决定使用表现最好的 Gainratio 方法作为属性评价标准，用 ranker 搜索方法来与本文提出的快速属性选择方法进行比较，在每个数据集上均作5次运算取其平均值计数。

表1 两种属性选择方法的结果

属性选择方法	原属性数	选择的属性数	记录数	属性名称
Gainratio	41	12	50000	root_shell\wrong_fragment\num_compromised diff_srv_rate\protocol_type\num_failed_logins\hot\flag\serror_rate\srv_serror_rate\service\logged_in
快速属性选择	41	9	50000	service\src_bytes\dst_bytes\logged_in\flag\dst_host_srv_count\dst_host_diff_srv_rate\dst_host_srv_diff_host_rate\dst_host_serror_rate

表2 属性选择性能比较

	属性选择及建模时间	属性选择及建模空间
Gainratio + C4.5	18.230.1s	23,108K10K
快速属型选择 Naïve Bayesian	9.5120.08s	15,388K12K

表3 分类结果

	C4.5	NBC
训练数据时间消耗	2.010.04s	1.100.07s
训练数据空间消耗	34,700K±10K	2802K± 8K
训练数据正确率	99.612%	99.98%
测试数据时间消耗	7.56±0.03s	4.30 ±0.08s
测试数据空间消耗	157,000±60K	3806K±12K
测试数据正确率	99.52%	99.55%

由于采用不同属性选择方法及不同的分类器，使得选择出的属性也不尽相同，这点也说明 不同的分类器适合不同的训练属性。

从实验结果可以看出，由于本文提出的快速属型选择算法与 Gainratio 算法均需在内存中保留原始数据进行后续计算与建立分类模型，因而牺牲了一定的空间性能来节省时间上的开销。这种空间的消耗也是在可以接受的范围，且本文提出的快速属性选择算法在时、空性能上均较 Gainratio 有一定提高，分别提升了47.8%及33.4%。

在结合各自属性选择方法后的入侵检测分类结果上，快速属型选择算法与朴素贝叶斯分类器有更出色的表现。由于C4.5算法需要在内存中保存全部数据以方便计算，且构建的决策树较庞大，而本文提出的方法可以连续读入数据再进行处理，分类模型仅在内存中保留一种表的结构，因而本文方法的内存消耗在两个数据集上分别仅有C4.5算法内存消耗的1/12和1/42。由于朴素贝叶斯的条件独立假设简化了计算，使得其在分类时间上有一定优势。C4.5的高内存消耗是由于其采用了牺牲空间以换取时间的方法，但是我们注意到这样的做法并没有使得其分类时间优于朴素贝叶斯分类器，两者的分类精度都是令人满意的。

总结与展望 通过以上实验及分析，本文提出的快速属性选择方法可以高效地从大规模入侵检测数据中获得适合朴素贝叶斯分类器的属性。随着训练数据量的增加，训练模型精度进一步提升且时间及空间开销仅为线性增加，因而具有较好的实用性。当测试数据同训练数据的概率分布出现较大变动时，朴素贝叶斯分类器的准确率会受到严重影响，因而将朴素贝叶斯分类器应用于真实网络环境时需要不断地去探测网络数据分布是否发生了重大的改变，进而更新分类模型以适应这种改变。由于快速属性选择可以高效地选择代表当前数据特征的属性，使得贝叶斯分类模型的更新代价进一步降低。在今后的工作中，将找寻如何度量当前网络数据是否发生改变的方法，进一步增强贝叶斯分类模型的适应性。

参考文献

1 Han J W, Kamber M. Data Mining: Concepts and Techniques. San Francisco, CA: Morgan Kaufmann, 2000
2 Ayoade J. Feature deduction and ensemble design of intrusion detection systems. Article Computers & Security, 2005,24(6)
3 史忠植. 知识发现[M]. 北京：清华大学出版社，2002
4 Mitchell T M. 机器学习[M]. 曾华军，张银奎，等译. 北京：机械工业出版社，2003
5 王国胤，于洪，杨大春. 基于条件信息熵的决策表约简[J]. 计算机学报，2002，25(7)：759～766
6 Liu Huan, Setiono R. Feature selection and classification—a probabilistic wrapper approach[C]. In: Proceedings of the Ninth International Conference on Industrial and Engineering Applications of AI and ES
7 翟素兰，郑诚. 用于入侵检测的基于粗糙集的贝叶斯分类器. 计算机技术与发展，2006，01-0226-02
8 Information and Computer Science University of California. Irving KDD cup 1999 Data [EB/OL]. http://kdd.ics.uei.edu/databases/kddcup99/kddcup99.html,
9 胡学钢，郭亚光. 一种基于粗糙集的朴素贝叶斯分类算法. 合肥工业大学学报(自然科学版)，2006(2)

(上接第141页)

索时，用“forest”标注的图像也作为满足条件的图像返回，此时返回的图像数目为71。

如果已标注的图像库中存在用查询的关键词标注的图像，而与该查询关键词同属一个同义词集合的其它单词没有用于图像标注，则用该关键词检索时，基于概念的图像检索和基于关键词的图像检索，得到的图像数量是相同的。如检索词“wall”用于基于关键词和基于概念的检索时，由于不存在“wall”同义词标注的图像，因此满足检索条件返回的图像数量均为98。

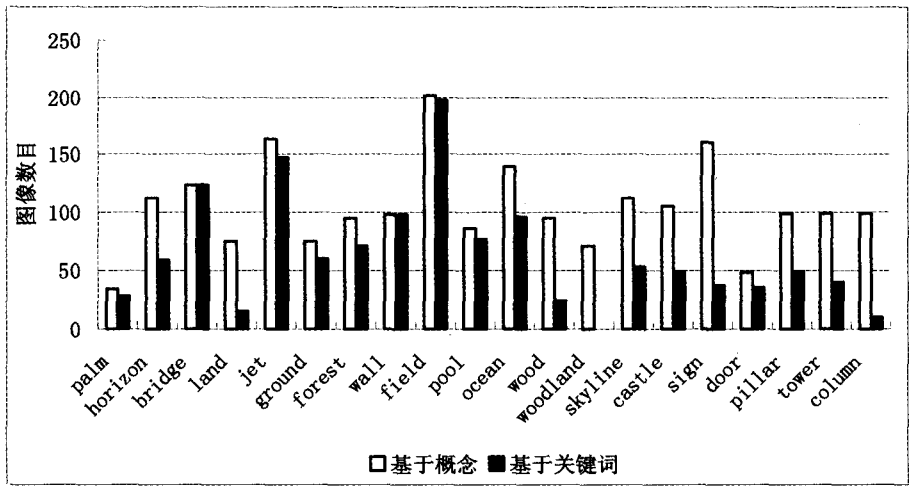

图2 实验结果

结论与展望 本文通过采用WordNet同义词集合，将图像检索所用的关键词扩展到概念层次的方法，实现了基于概念语义的图像检索，能部分解决因用户的理解不同而产生的对图像语义理解的歧义问题，从实验结果可以看出，能较好地提高检索的性能。下一步的工作考虑在概念关系的定义中引入单词间的“上下位关系”、“部分与整体的关系”等复杂的语义关系，进一步提高图像检索的性能。

参考文献

1 Smeulders A, Worring M, Santini S, et al. Content-based image retrieval at the end of the early years. IEEE Trans Pattern Analysis and Machine Intelligence, 2000,22(12):1349～1380
2 Flickner M, et al. Query by image and video content. The QBIC system. IEEE Computer, 1995, 28 (9): 23～32
3 Bach J R, Fuller C, Gupta A, et al. Virage image search engine: an open framework for image management. SPIE Storage and Re2 trieval of Image and Video DataBases, 1996, 4: 76～874
4 Smith J, Chang S F. VisualSEEK: A fully automated content-based image query system. In: Proceedings of the 4th ACM Multimedia Conference, Boston MA, USA, 1996. 87～98
5 Rui Y, Thomas S H, Chang S F. Image retrieval: Past, present and future. Journal of Visual Communication and Image Representation, 1999, 10 (1): 39～62
6 Woods W. Conceptual Indexing: a better way to organize knowledge: [Technical Report, SMLI TR-97-61]. Sun Microsystems Laboratories, Mountain View, USA, 1997
7 Clark P, Thompson J, Holmback H, et al. Exploiting a Thesaurus-based Semantic Net for Knowledge-based Search. In: Proceedings of the Twelfth Conference on Innovative Applications of AI (AAAI/IAAI'00), 2000. 988～995
8 Web Image Learning for Searching Semantic Concepts in Image Databases. WWW2004, New York, USA, ACM, May 2004 1-58113-912-8/04/0005
9 Cheng P J, Chien L F. Effective Image Annotation for Search Using Multi-level Semantics. Journal of Digital Libraries: Special Issue on Asian Digital Libraries, 2004. 258～271
10 Duygulu P, Barnard K, de Freitas N, et al. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In: the Proceeding of Seventh European Conference on Computer Vision (ECCV2002), Copenhagen, Denmark, May - June 2002