

基于多特征选取和类完全加权的入侵检测

李 蓉,周维柏

(华南师范大学增城学院,广东 广州 511363)

摘要:为提升入侵检测系统的整体性能,文中提出一种新的算法。首先使用孤立点滤除算法进行数据前期处理,通过特征选取算法筛选出各分类器中最佳的特征空间,以增强各分类算法的训练模型。再进一步运用十倍交叉验证法对分类模型实施性能评估,把具有最佳捕捉率的分类模型作为预测测试样本类别时的加权分类模型,最后得出整体推论结果。仿真实验表明该算法整体分类准确率提高到96%,成本值减低为0.1983,能够成功地改善网络异常入侵检测的分类性能。

关键词:入侵检测;数据挖掘;孤立点检测;多特征选取;类完全加权

中图分类号:TP393.08

文献标识码:A

文章编号:1673-629X(2014)07-0145-04

doi:10.3969/j.issn.1673-629X.2014.07.036

Intrusion Detection Based on Multiple Feature Selection and Class Fully Weighted

LI Rong,ZHOU Wei-bai

(Zengcheng College of South China Normal University,Guangzhou 511363,China)

Abstract:In order to improve the performance of intrusion detection system,a new algorithm is proposed. Firstly,the outlier deletion algorithm is used to obtain the training data in the data preprocessing phase. Secondly,the multiple feature selection algorithm is used to find out the best feature space for the classifiers,and then the training models of the classifiers could be well trained. Furthermore,the ten fold-cross validation is applied to evaluate the performances of the classification models,and the classification models with best recalls are used as the weighted classification models in the class fully weighted algorithm to predict the classes of test data. Finally,the inference results are concluded. Simulation results show that the classification accuracy of this algorithm reaches 96%,the cost value is 0.1983,can enhance performance of the network intrusion detection system.

Key words:intrusion detection;data mining;outlier detection;multiple feature selection;class fully weighted

0 引言

利用数据挖掘技术构建入侵检测系统,通过其学习算法从网络数据流量中自动学习分类规则,建立分类模型来识别正常或异常行为,并产生新的规则来侦测未知的入侵行为,这已成为近年来应用于入侵检测系统最广泛的技术^[1-2]。

数据挖掘技术可将网络数据流量分类,通过统计和机器学习算法,启发性地提取出隐藏且具有价值的知识与规律,进一步达到预测与支援决策的功能,主要步骤为:数据前期处理、数据属性选择、分类模块产生^[3-4]。但由于数据挖掘技术本身的一些特性,致使应用到检测系统后存在较高的误报率,数据挖掘时间长,很难达到实时。因此提出一种有效的算法机制,改

善孤立点对数据集的影响、有效的特征选取机制及整合专家系统优化分类能力,以提升网络入侵检测系统的分类能力和整体性能。

1 基于多特征选择和类完全加权的入侵检测

文中提出孤立点滤除算法来滤除数据集中孤立点,以获得具有代表性训练样本,然后使用多样化特征选取算法来降低无关特征的影响,以强化特定分类器的分类效果,最后使用类完全加权算法来选定类别的加权分类模型,作为预测类别判断的决策基准,逐步强化分类模型,提升入侵检测系统的性能。算法的模型如图1所示。

收稿日期:2013-09-16

修回日期:2013-12-25

网络出版时间:2014-04-24

基金项目:广东省自然科学基金(S2011010003442)

作者简介:李 蓉(1972-),女,广东广州人,讲师,硕士,研究方向为模式识别、网络安全。

网络出版地址:<http://www.cnki.net/kcms/detail/61.1450.TP.20140424.0818.061.html>

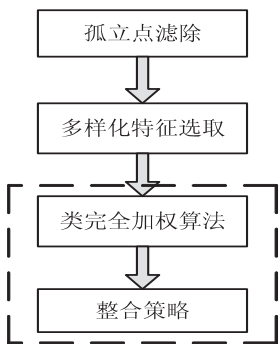


图 1 算法的模型

1.1 孤立点滤除算法

入侵检测系统数据库中通常包含大量的网络数据,这些数据中大部分对建立分类模型有意义,但也存在孤立点,这些孤立点对整体分类效率有非常大的影响,当少数类别的样本稀少时,少量的孤立点将足以误导分类算法训练模型的建立,无法建立准确的分类规则,从而影响整体分类效果^[5-6]。若使用单一特定算法来评估样本是否为孤立点,可能会受到特定算法在训练过程中进行推论与归纳时的归纳偏移影响,误将正常样本同孤立点一起删除;但若不去除数据中的孤立点,所建立训练模型对测试样本的分类效果将难以提升。因此文中采用以统计孤立点为基础,使用十倍交叉验证法的方法,同时结合多数投票与交叉验证机制,实施训练样本中孤立点的挑选。算法步骤如下:

①基本分类器训练模型建立:由 Decision Tree、Adaboost、Bagging、Random Forest 等基本分类器,运用十倍交叉验证法针对训练样本集来建立训练模型。

②分类器设定:在 Decision Tree、Adaboost、Bagging、Random Forest 等基本分类器中选择一种,作基准分类器,其余 3 种分类器作为对照组分类器。

③多数投票:在对照组分类器判断为错误的样本中,进行错误样本多数投票,若达三分之二以上的分类器判定为错误样本,则归类为候选孤立点。

④比对挑选孤立点:将候选孤立点与基准分类器判断为错误样本者进行交叉比对,若为共同错误样本,则视为孤立点,并从原来的训练样本集中滤除。

⑤判定训练终止条件:重复步骤①至步骤④,直到每个基本分类器都已设定为基准分类器为止。

孤立点滤除算法首先通过较佳的基本分类器个别进行十倍交叉验证法,筛选出正确与错误的分类结果;再经过多数投票与交叉对比的机制,来过滤训练样本集中的孤立点,进而获得训练样本集中具有代表性的训练样本,作为后续强化基本分类器训练模型之用。

1.2 多样化特征选取

一般情况下,入侵检测系统需处理大量的数据,会消耗相当多的资源,数据中包含不相关或冗余的特征,

在训练的过程中损耗比较多的资源,造成较长的训练时间及较差的检测性能^[7-8]。所以去除这些冗余特征或不重要特征是非常重要的。为解决上述问题并建立更为准确的分类模型,特征选取是一种有效的方式。

特征选取技术可分为包装(Wrapper)与滤波器两种。滤波器方法主要是分别去计算每个数据样本的特征权重值,持续加入或删除特征,并利用学习算法中的目标函数评估特征集合的优劣。常用的有 Information Gain、Relief、Gain ratio 等,不同的特征选取算法所计算出的特征权重值重新排序后将不同,若能选择适当且不同的分类器来搭配不同的重要特征空间,即可训练出具有差异性的分类模型,将有助于专家系统整体分类性能的改进。故文中利用 Information Gain、Relief、Gain ratio 特征选取方法,配合 Decision Tree、Adaboost、Bagging、Random Forest 分类器,提出多样化特征选取算法。具体步骤如下:

①权重计算:轮流应用 Information Gain、Relief、Gain ratio 特征选取算法,计算出该训练样本集所有特征值的权重。

②特征子集合选定:依特征值的权重作排序,并依序选取一项特征值加入特征子集合中。

③分类模型训练:轮流应用 Decision Tree、Adaboost、Bagging、Random Forest 等基本分类器,对步骤②所选出的特征子集合,使用 Information Gain、Relief、Gain ratio 特征选取算法,建立不同特征空间的分类模型。

④判定训练终止条件:重复步骤②至步骤③,直到所有特征值均已选定完毕。

1.3 类完全加权算法

由于不同的分类模型对于特定的类别有较好的分类效果^[9-13],在专家系统对类别预测时,若能对各分类模型赋予不同的权重值,并在最后推理时根据此权重值来实施决策,应可获得更好的分类效果。对多样化特征选取算法产生的分类模型,利用十倍交叉验证法进行性能评估,并记录各分类模型对每一类的捕捉率及整体准确率,最后在十倍交叉验证法中,把具有最好捕捉率的分类模型作为预测测试样本类别时的加权分类模型。当加权分类模型判断结果相互冲突时,则依前阶段十倍交叉验证后的整体准确率较高者为最终的加权分类模型。算法的具体步骤如下:

①训练模型选定:选定经孤立点滤除算法和多样化特征选取算法后,其整体准确率优于仅实施孤立点滤除算法的分类模型为训练模型。

②训练模型验证与评估:各训练模型实施十倍交叉验证,并记录各分类模型对每一类别的捕捉率及整体准确率。

③加权分类模型选定:将每一类选定具有最好捕捉率的分类模型作为预测测试样本类别时的加权分类模型。

④类别决策基准:对测试样本利用各加权分类模型进行类别预测,类别的决定以该类别的加权分类模型的决策为基准;当各加权分类模型类别的判断结果相互冲突时,则依步骤②中具有最高整体准确率的加权分类模型的决策为基准。

1.4 整合策略

主要是用以结合专家系统内不同分类模型输出推理结果。

2 实验结果与分析

文中研究以 Weka 和 Java 来构建实验环境,数据集采用 KDD99。

2.1 数据集描述

该研究采用 KDD99 数据集,原始的测试数据中每个连接包含 41 种定性和定量的特征,并标识为正常或攻击,这些特征属性分为离散和连续两种属性。实验采用数据集的 10% 训练样本,实验样本数量和分布情况如表 1。

表 1 实验数据数量及分布情况

攻击类型	训练数据集	攻击数据集
Normal	972 78	60 593
DoS	391 458	229 853
Probe	4 107	4 166
U2R	52	228
R2L	1 126	16 189

2.2 检测算法性能评估方式

为了评估分类模型的性能,文中定义准确率和错误率,分别如式(1)、式(2):

准确率=
$$\frac{TP+TN}{TP+FP+FN+TN}$$
(1)

错误率=
$$\frac{FP+FN}{TP+FP+FN+TN}$$
(2)

其中,TP 指使用分类模型将原本属于正类的样本正确预测为正类;FP 指使用分类模型将原本属于反类的样本错误预测为正类;TN 指使用分类模型将原本属于反类的样本正确预测为反类;FN 指使用分类模型将原本属于正类的样本错误预测为反类。

除准确率和错误率外,还定义如下衡量指标:

Recall=
$$\frac{TP}{TP+FN}$$
(3)

Precision=
$$\frac{TP}{TP+FP}$$
(4)

F-measure=
$$\frac{2\times\text{Recall}\times\text{Precision}}{\text{Recall}+\text{Precision}}$$
(5)

其中,捕捉率(Recall)越高代表正类被正确分类的比例越高;精确率(Precision)越高则代表分类模型越少将原本属于反类的样本错误预测为正类;F-measure 则是结合捕捉率与精确率的综合评价指标。

2.3 实验结果

实验分三部分:第一部分,对运用孤立点滤除算法进行训练样本孤立点去除实验;第二部分是运用多样化特征选取的性能比较;第三部分是算法性能测试。

2.3.1 孤立点滤除算法性能分析

该实验分类算法采用 Weka 内建的 Decision Tree、Adaboost、Bagging、Random Forest 等集成分类算法,孤立点滤除前后整体分类准确率比较如表 2。

表 2 孤立点滤除前后整体分类准确率比较 %

	Decision Tree	Adaboost	Bagging	Random Forest
滤除前	92.230 0	90.111 2	92.379 5	92.350 9
滤除后	92.606 5	90.111 2	92.649 9	92.518 4
差异	0.376 5	0	0.270 4	0.167 5

经孤立点滤除后仅 Adaboost 分类器所得的分类效果不变,其余三个都有提升,改善幅度以 Decision Tree 最大。从实验结果可看出,利用孤立点滤除算法进行孤立点滤除,可改善及强化分类性能。

2.3.2 多样化特征选取算法性能分析

针对实验一滤除孤立点后更新的训练样本,分别利用 Information Gain、ReliefF、Gain ratio 三种特征选取机制,计算出所有特征值的权重,并以特征值的权重作排序,训练特征权重排序结果如表 3。

表 3 不同特征选取方法的特征排序

选取方法	特征排序
ReliefF	3,36,2,12,33,23,4,34,29,35,24,38,30,39,25,26,14,37,13,8,40,31,10,17,27,41,22,28,11,1,16,6,18,19,5,7,20,21,9,15
	12,11,6,14,22,9,37,3,32,31,5,2,1,17,23,36,18,19,16,10,15,24,38,35,25,33,39,34,30,26,4,41,40,29,27,13,28,8,21,7,20
Gain ratio	5,23,3,6,12,24,36,32,2,37,33,35,34,31,30,29,38,39,25,4,26,1,40,41,27,28,10,22,16,19,13,17,11,8,14,18,9,15,21,7,20
InfoGain	

依序选取前 1 项至 41 项的特征子集合,利用 Decision Tree、Adaboost、Bagging、Random Forest 四种分类器,配合使用三种特征选取方法选出的特征子集合,进行分类模型的训练,结合孤立点滤除和多样化特征选取,建立具有不同特征的特征空间和基本分类器。

同表 2 比较,使用多样化特征选取算法后准确率都有提升。

2.3.3 类别完全加权算法性能分析

对于采用多数投票整合策略的专家系统,分类模型采用孤立点滤除和多样化特征选取后所建立具有最佳特征空间的训练模型(DT-RF-30、Bag-GR-20、RF-

GR-15)。在训练模型评估与验证阶段,对所选定的训练模型实施十倍交叉验证法,并记录各模型对每一个类别捕捉率和整体准确率,如表 4。

表 4 十倍交叉验证法捕捉率及整体准确率结果 %

	DT-RF-30	Bag-GR-20	RF-GR-15
Normal	94.2	93.5	92.5
DoS	100	99.9	99.9
Probe	98.4	95.9	96
U2R	82.5	90.4	89.5
R2L	67	87.5	85.1

该算法与使用传统专家系统的入侵检测系统性能比较如表 5。

表 5 该算法与使用传统专家系统的入侵检测系统性能比较

类别	评估指标	传统专家系统	文中算法
Normal	Recall	99.35%	99.52%
	F-measure	86.63%	85.34%
DoS	Recall	97.4%	97.43%
	F-measure	98.11%	98.64%
Probe	Recall	78.23%	79.22%
	F-measure	84.1%	86.15%
U2R	Recall	10.53%	16.67%
	F-measure	18.6%	27.54%
R2L	Recall	12.56%	13.72%
	F-measure	21.95%	24.73%
Accuracy		93.068%	96.127%
Cost		0.213	0.208 3

由表 5 可以看出,文中算法的整体分类准确率提升到 96.127%,而成本降低为 0.208 3。

3 结束语

文中算法提出孤立点滤除、多特征选取、类别完全加权算法来提升入侵检测系统的检测性能,实验表明该算法可成功改善网络异常入侵检测的分类效能。未来若能通过赋予各分类模型不同的成本值,在最后推

论时实施权重投票,应可获得最佳的分类效果。

参考文献:

[1] Lu Huibin,Xu Gang. A new intrusion detection method based on data mining[J]. Microprocessors,2006,27(4):58-60.

[2] Li Hanguang,Ni Yu. Intrusion detection technology research based on apriori algorithm[C]//Proc of 2012 international conference on applied physics and industrial engineering. Hong Kong:[s. n.],2012:1615-1620.

[3] Wu Suyun,Yen E. Data mining-based intrusion detectors[J]. Expert Systems with Applications,2009,36(3):5605-5612.

[4] 李 睿,肖维民. 基于孤立点挖掘的异常检测研究[J]. 计算机技术与发展,2009,19(6):168-170.

[5] 罗 敏,阴晓光,张焕国,等. 基于孤立点检测的入侵检测方法研究[J]. 计算机工程与应用,2007,43(13):146-149.

[6] 黄 斌,史 亮,姜青山,等. 基于孤立点挖掘的入侵检测技术[J]. 计算机工程,2008,34(3):88-90.

[7] Kamal A H M,Zhu Xingquan,Pandya A,et al. Feature Selection with biased sample distributions[C]//Proceedings of the IEEE international conference on information reuse and integration. Las Vegas,NV:IEEE,2009:23-28.

[8] Kamal A H M,Zhu Xingquan,Pandya A S,et al. Feature selection for datasets with imbalanced class distributions[J]. International Journal of Software Engineering and Knowledge Engineering,2010,20(2):113-137.

[9] 赵晓峰,叶 震. 基于加权多随机决策树的入侵检测模型[J]. 计算机应用,2007,27(5):1041-1043.

[10] 王鹏英,黄 海,黄晓平. 基于加权特征筛选的入侵检测系统[J]. 计算机科学,2012,39(1):89-91.

[11] Tsai Chih-Fong,Hsu Yu-Feng,Lin Chia-Ying,et al. Intrusion detection by machine learning:a review[J]. Expert Systems with Applications,2009,36(10):11994-12000.

[12] 王 骐,王青萍. 一种基于特征的入侵检测模块的优化布置算法[J]. 计算机仿真,2011,28(6):136-140.

[13] 夏永祥,史志才. 基于 GPU 和特征选择的 SVM 入侵检测模型[J]. 计算机工程,2012,38(8):111-113.

术与发展,2010,20(4):17-20.

[13] Spiliopoulou M,Mobasher B,Berendt B,et al. A framework for the evaluation of session reconstruction heuristics in Web-usage analysis[J]. INFORMS Journal on Computing,2003,15(2):171-172.

[14] Stevanovic D,Vlajic N,An A. Detection of malicious and non-malicious Website visitors using unsupervised neural network learning[J]. Applied Soft Computing,2013,13(1):698-708.

[15] Han Jiawei,Kamber M. 数据挖掘概念与技术[M]. 范 明,孟小峰,译. 北京:机械工业出版社,2012.

(上接第 144 页)

[8] 刘加伶,范 军. 基于用户访问树的 Web 日志挖掘数据预处理[J]. 计算机科学,2009,36(9):154-156.

[9] 李 志. 基于 Web 服务器日志挖掘的数据预处理技术研究[D]. 成都:电子科技大学,2012.

[10] 顾兆军,李晓红,王 伟,等. Web 日志挖掘中的会话识别方法研究[J]. 计算机技术与发展,2012,22(4):45-49.


[11] 杨 楠. 基于关联规则 Apriori 算法的 Web 日志挖掘[D]. 成都:成都理工大学,2012.

[12] 方 杰,朱京红. 日志挖掘中的数据预处理[J]. 计算机技

基于多特征选取和类完全加权的入侵检测

作者：[李蓉](#)，[周维柏](#)，[LI Rong](#)，[ZHOU Wei-bai](#)

作者单位：[华南师范大学增城学院, 广东 广州, 511363](#)

刊名：[计算机技术与发展](#)

英文刊名：[Computer Technology and Development](#)

年，卷(期)：2014(7)

参考文献(13条)

1.[Lu Huibin;Xu Gang](#) [A new intrusion detection method based on data mining](#) 2006(04)

2.[Li Hanguang;Ni Yu](#) [Intrusion detection technology research based on apriori algorithm](#) 2012

3.[Wu Suyun;Yen E](#) [Data mining-based intrusion detectors](#) 2009(03)

4.[李睿;肖维民](#) [基于孤立点挖掘的异常检测研究](#) 2009(06)

5.[罗敏;阴晓光;张焕国](#) [基于孤立点检测的入侵检测方法研究](#) 2007(13)

6.[黄斌;史亮;姜青山](#) [基于孤立点挖掘的入侵检测技术](#) 2008(03)

7.[Kamal A H M;Zhu Xingquan;Pandya A](#) [Feature Selec-tion with biased sample distributions](#) 2009

8.[Kamal A H M;Zhu Xingquan;Pandya A S](#) [Feature se-lection for datasets with imbalanced class distributions](#) 2010(02)

9.[赵晓峰;叶震](#) [基于加权多随机决策树的入侵检测模型](#) 2007(05)

10.[王鹏英;黄海;黄晓平](#) [基于加权特征筛选的入侵检测系统](#) 2012(01)

11.[Tsai Chih-Fong;Hsu Yu-Feng;Lin Chia-Ying](#) [Intru-sion detection by machine learning:a review](#) 2009(10)

12.[王骐;王青萍](#) [一种基于特征的入侵检测模块的优化布置算法](#) 2011(06)

13.[夏永祥;史志才](#) [基于GPU和特征选择的SVM入侵检测模型](#) 2012(08)

本文链接：http://d.wanfangdata.com.cn/Periodical_wjfz201407036.aspx