K-MEANS算法在IDS中的应用研究

李玲娟,李 冰,薛 明
(南京邮电大学 计算机学院,江苏 南京 210003)

摘 要:聚类算法广泛应用于入侵检测系统(IDS)的数据挖掘中。虽然K-MEANS算法是最为经典的聚类算法之一,但是由于入侵检测系统的数据集具有特殊性,直接在其上进行K-MEANS聚类的效果不佳。为了提高K-MEANS在IDS数据集上的聚类准确性,引入一种数据预处理方法。该方法对IDS的记录特征做标准化处理,使原本取值范围差异很大的数值型特征在同一个区间内取值,排除原始数据中不同度量带来的不良影响,从而优化聚类的效果。仿真实验表明,K-MEANS算法对预处理后的IDS数据集的聚类准确度有很大的提高。

关键词:数据挖掘;入侵检测系统;K均值聚类;预处理
中图分类号:TP311 **文献标识码**:A **文章编号**:1673-629X(2010)07-0129-03

Research on Application of K-MEANS Algorithm in IDS

LI Ling-juan, LI Bing, XUE Ming
(College of Computer, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

Abstract: Clustering algorithms are widely used in intrusion detection system (IDS) to mine the data. Although K-MEANS is one of the most classical clustering algorithms, the effect is not very good when it is used in IDS directly. The reason is that the data set of intrusion detection system is peculiar. In order to improve the clustering accuracy of K-MEANS on IDS data set, designs a data preprocessing method, which makes the features of IDS record standardized, and makes all features with very different value ranged in the same range. This can exclude the impact of difference between the measured variables of the original data, and can help to improve the effect of clustering. Simulation results show that the clustering accuracy of K-MEANS on the preprocessed IDS data set has been greatly improved.

Key words: data mining; intrusion detection system; K-MEANS clustering; preprocessing

0 引 言

入侵检测系统(Intrusion Detection System, IDS)是用来发现外部攻击和合法用户滥用特权的一种方式,是动态安全技术中最核心的技术之一[1]。它从系统内部和各种网络资源中主动采集信息,分析可能的入侵攻击行为。根据数据分析手段的不同,常用的入侵检测方法分为误用检测和异常检测[2]。误用检测是利用已知的入侵方法进行入侵活动的检测,异常检测是利用定量的方式来描述可接受的行为特征,以区分和正常行为相违背的异常行为特征来检测入侵。

近年来为了提高IDS的检测效率和智能性,数据挖掘技术正被越来越广泛地应用于IDS中,它是一种从大量数据中提取人们感兴趣的、事先未知的知识和规律的技术[3,4]。数据挖掘的聚类分析常被用于IDS的检测阶段。所谓聚类是一种无监督的学习过程[5,6],其目标是基于一定的数据度量标准,把给定的d维空间中的n个数据点聚集成c个类,使类内数据点的相似度极大化,类间数据点的相似度极小化。

K-MEANS是聚类的经典算法之一,此算法容易实现,而且时间和空间复杂度相对较小。但是由于IDS中被检测数据集的各个特征的取值差异比较大[7],直接用K-MEANS进行聚类的效果不佳。为此,文中引入了一种数据预处理算法,使得K-MEANS算法能在预处理后的IDS数据集上获得好的聚类效果,进而提高检测效率。

1 K-MEANS算法概述

K-MEANS算法以k为输入参数,把n个对象的集合分为k个簇,使得结果簇内的相似度高,而簇间的相似度低[6]。

K-MEANS算法可描述如下[6]:

算法:K-MEANS。

收稿日期:2009-10-14;**修回日期**:2010-01-21
基金项目:国家自然科学基金(60863001);江苏省高校自然科学基础研究项目(08KJB620002);南京邮电大学校科研基金(NY207051)
作者简介:李玲娟(1963-),女,辽宁辽阳人,教授,研究方向为数据挖掘、网络安全等。

输入:簇的数目 k 和包含 n 个对象的数据集。

输出:k 个簇的集合。

步骤:

1) 任意选择 k 个对象作为初始的簇中心;

2)repeat

3)根据簇中对象的平均值,将每个对象(重新)赋给最相似的簇;

4)更新簇的均值;

5)until 不再发生变化

K-MEANS 算法的处理流程[6]是:首先,随机地选择 k 个对象,每个对象初始地代表了一个簇的平均值或中心。对剩余的每个对象,根据其与各个簇中心的距离,将它赋给最近的簇。然后重新计算每个簇的平均值。这个过程不断重复,直到准则函数收敛[6]。通常,采用平方误差准则,其定义如下:

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} | p - m_i |^2 \tag{1}$$

这里 E 是数据库中所有对象的平方误差的总和。p 是空间中的点,表示给定的数据对象。m_i 是簇 C_i 的平均值(p 和 m_i 都是多维的)。这个准则试图使生成的结果簇尽可能地紧凑和独立[6]。

2 IDS 数据源特点分析

可以借助 MIT 林肯实验室的 KDD99 数据集[8]来分析 IDS 的数据源,此数据集是在军事网络环境中运用非常广泛的模拟入侵攻击所得到的网络数据集,包括近 500 万条通过对 TCP 数据帧进行预处理得到的网络连接记录。其中的每条记录由 42 个特征组成,前面 41 个特征涉及 TCP 连接的基本特征、内容特征、流量特征等,其中有 38 个数值型特征,3 个字符型特征,最后一个特征反映该记录属正常行为还是某种攻击[9]。不同特征的取值多少和取值范围差异都很大,比如 duration 有 2495 个值、src_bytes 有 3300 个值,而 land 只有 2 个值、is_hot_login 只有 1 个值[9]。

数据集中的攻击类型分别属于四大类[10]:拒绝服务类攻击 DOS,比如 smurf 攻击;远程系统未授权访问类攻击 R2L,比如 warezclient 攻击;未授权或非法使用超级用户权限类攻击 U2R,比如 rootkit 攻击;监视或刺探系统脆弱性的攻击 Probing,比如 portsweep 攻击。

3 数据预处理算法

考虑到 IDS 中被检测记录的特征取值特点会影响 K-MEANS 的聚类效果,文中基于经典的数学定义[11],设计了适用于 K-MEANS 在 IDS 中聚类的数据预处理方法,其目的是使各特征在同一个区间内取值。具体方法如下:

计算平均绝对偏差:

$$s_f = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_{if} - m_f)^2} \tag{2}$$

其中:$X_{1f}, \cdots, X_{nf}$ 是特征 f 的 n 个特征值,m_f 是 f 的平均值,即:

$$m_f = \frac{1}{n} \sum_{i=1}^{n} X_{if} \tag{3}$$

计算标准化特征值:

$$Z_{if} = \frac{X_{if} - m_f}{s_f} \tag{4}$$

通过上述标准化过程可以使特征的平均值变为 0,标准差变为 1。最后由公式(5) 产生归一化结果:

$$Z'_{if} = \frac{Z_{if}}{\max\limits_{1 \leqslant i \leqslant n} \{Z_{if}\} - \min\limits_{1 \leqslant i \leqslant n} \{Z_{if}\}}$$
$$\max_{1 \leqslant i \leqslant n} \{Z'_{if}\} - \min_{1 \leqslant i \leqslant n} \{Z'_{if}\} = 1 \tag{5}$$

4 仿真实验及结果分析

4.1 实验数据源简介

为了验证所引入的预处理方法的有效性,文中基于 KDD Cup 99 数据集做了仿真实验。

本仿真实验选取了 1 万条记录来比较数据预处理前后 K-MEANS 算法的聚类效果,其中包含 normal 记录和 4 种攻击记录各 2000 条,详见表 1。

表 1 实验数据的选择

记录类型	数量(条)
normal.	2000
smurf.	2000
rootkit.	2000
warezclient.	2000
portsweep.	2000

4.2 数据预处理算法的实现

文中用 C 语言编写的数据预处理代码如下:

```
//计算每个特征的平均值,放入数组 M 中
double M[41] = {0};
for(j=0;j<41;j++)
{
for(i=0;i<Cont;i++)
{
M[j] += K[i][j]/Cont;
}
}
//计算平均的绝对偏差值,放入数组 P 中
double P[41] = {0};
for(j=0;j<41;j++)
{
for(i=0;i<Cont;i++)
{
```

```
P[j] + = (K[i][j] - M[j]) * (K[i][j] - M[j])/(Cont - 1);
}
P[j] = sqrt(P[j]);
}
//计算标准化的特征值,放入二维数组 Q 中
for(j = 0;j < 41;j + + )
{
for(i = 0;i < Cont;i + + )
{
if (P[j] = = 0) Q[i][j] = 0;
else
Q[i][j] = (K[i][j] - M[j])/P[j];
}
}
//求出标准化特征值中的最大值与最小值,分别存入数组 Max 和 Min 中,同时求出(Max[j] - Min[j]),结果存入数组 N 中
double Max[41] = {0},Min[41] = {0},N[41] = {0};
for(j = 0;j < 41;j + + )
{
for(i = 0;i < Cont;i + + )
{
if (Max[j] < = Q[i][j]) Max[j] = Q[i][j];
if (Min[j] > = Q[i][j]) Min[j] = Q[i][j];
}
N[j] = Max[j] - Min[j];
}
//产生归一化结果,放入二维数组 R 中
for(j = 0;j < 41;j + + )
{
for(i = 0;i < Cont;i + + )
{
if(N[j] = = 0)
R[i][j] = 0;
else
R[i][j] = Q[i][j]/N[j];
}
}
```

4.3 实验结果与分析

实验中对选出的 1 万条记录运行以上代码进行预处理,并分别对原始的 1 万条记录和预处理后的 1 万条记录用 K-MEANS 算法(实现代码略)进行聚类,实验中取 $k = 5$,即聚类成 5 个簇,分别命名为 a、b、c、d、e。实验结果如表 2 所示。

由表 2 不难看出,对数据集直接运用 K-MEANS 算法来进行聚类,效果比较差,基本没有达到聚类的效果。用文中的预处理方法对实验数据进行预处理后,聚类效果明显改善。smurf 攻击、warezclient 攻击、portsweep 都被很好地聚类,normal 次之,只有 rootkit 攻击的聚类效果有限。

表 2　仿真结果对比

簇别	对应攻击类型	无预处理的 K-MEANS 聚类		预处理后的 K-MEANS 聚类	
		数量	簇内百分比	数量	簇内百分比
a	normal	0	0	0	0
	smurf	0	0	2000	100%
	rootkit	0	0	0	0
	warezclient	120	98.36%	0	0
	portsweep	2	1.64%	0	0
b	normal	10	4.50%	0	0
	smurf	0	0	0	0
	rootkit	0	0	0	0
	warezclient	44	19.82%	2	0.10%
	portsweep	168	75.68%	1936	99.90%
c	normal	730	42.94%	1580	100%
	smurf	0	0	0	0
	rootkit	400	23.53%	0	0
	warezclient	570	33.53%	0	0
	portsweep	0	0	0	0
d	normal	1010	13.82%	50	2.61%
	smurf	2000	27.38%	0	0
	rootkit	1200	16.42%	600	31.35%
	warezclient	1266	17.33%	1262	65.94%
	portsweep	1830	25.05%	2	0.1%
e	normal	250	38.46%	370	14.41%
	smurf	0	0	0	0
	rootkit	400	61.54%	1400	54.52%
	warezclient	0	0	736	28.66%
	portsweep	0	0	62	2.41%

5 结束语

文中的研究表明,由于 IDS 记录集中各个特征的取值差异较大,使得 K-MEANS 无法发挥效力,但是,采用合适的方法可以使 K-MEANS 之类的经典数据挖掘算法在 IDS 中取得良好的挖掘效果。

事实上能适合所有领域的数据挖掘算法是不存在的[12]。解决上述问题有两种思路:(1)改进数据挖掘的相关算法以适应 IDS 数据源的特点;(2)对 IDS 的数据源进行预处理,使已有的数据挖掘算法能被有效地利用。

文中合理地采用了第二种思路,针对 IDS 数据源的特点和 K-MEANS 在 IDS 中的应用,设计了数据预处理方法,有效排除原始数据中变量之间的不同度量对聚类的影响,使待聚类的数据分布在同一区间内,以相同的量级参与聚类。仿真实验结果验证了文中的研究思路的正确性。需要指出的是:要进一步提高 K-MEANS 算法在 IDS 中的聚类准确度,可以在字符型特征的处理、循环条件的设定、各特征对判别攻击的重要性(即权重)的估计等方面做进一步的研究。这正是下一步的研究目标。

参考文献:

[1] Anderson J P. Computer security thread monitoring and surveillance[R]. Fort Washington, PA: James P Anderson Co,

(下转封三)

加有意义的稀有类实例数目，构建分类器分类。实验表明 EPRC 算法能够有效识别稀有类，提高分类的性能。

2005 年笔者提出了利用基本显露模式 eEP (essential Emerging Patterns) 的两阶段分类稀有类方法 (即 CREEPTP)。基本显露模式 eEP 是一种特殊的 EP，在稠密的数据集中通常包含大量的 EP，实践表明使用大量 EP 分类并不是很有效。H. Fan 和 K. Ramamohanarao 提出使用一种特殊的 EP，eEP (essential Emerging Patterns) 进行分类[11]。eEP 是 EP 边界表示的左边界的子集[12]，具有很高的增长率。eEP 去掉了那些包含冗余信息和噪声的项集，其数量比 EP 少得多，并且具有很高的分类效率。CREEPTP 使用 eEP 作为分类的基础，并采用两个阶段来挖掘基本显露模式 eEP，第二个阶段作为第一个阶段的纠正来提高稀有类分类的精度。使用评价稀有类分类的两个常用尺度：召回率 (Recall) 和精度 (Precision)。实验结果表明该算法在分类稀有类的时候可以取得较好的召回率和精度。实验结果和几个经典的分类算法比较，在分类准确性上也达到了很好的性能。

5 结束语

文中针对稀有类分类问题进行了研究。文章给出了稀有类分类问题的概念，并分析了稀有类分类问题的特征，从而给出稀有类分类问题的难点所在。针对稀有类问题，给出探讨了评估标准以及当前分类稀有类的主要方法。

参考文献：

[1] Han J, Kanber 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，2001.

[2] 刘艳霞，职为梅，杨　亮. 稀有类分类问题研究[J]. 微型机与应用，2005(6)：54－56.

[3] Weiss G, Provost F. Learning when training data are costly: the effect of class distribution on tree induction[J]. J. Aritif. Intell. Res., 2003, 19: 315－354.

[4] Yan Min, Kamel M S, Wong A K C, et al. Cost－sensitive boosting for classification of imbalanced data[J]. Pattern Recognition, 2007(10): 3358－3378.

[5] Agarwal R, Joşhi M V. PNrule: A new Framework for Learning Classifier Models in Data Mining (A Case－Study in Network Intrusion Detection)[C]//Proc. of the First SIAM Conference on Data Mining. Chicago, USA: [s. n.], 2001.

[6] Alhammady H, Ramamohanarao K. The Application of Emerging Patterns for Improving the Quality of Rare－class Classification[C]//Proc. of the 8th Pacific－Asia Conf. on Advances in Knowledge Discovery and Data Mining (PAKDD2004). Sydney, Australia: [s. n.], 2004: 207－211.

[7] 职为梅，范　明. 利用基本显露模式两阶段分类稀有类[J]. 微机发展(现更名：计算机技术与发展)，2005，15(12)：44－47.

[8] Agarwal R, Joshi M V, Kumar V. Mining Needles in a Haystack: Classifying Rare Classes via Two－Phase Rule Induction[C]//the Proc of ACM SIGMOD/PODS. [s. l.]: [s. n.], 2001: 91－102.

[9] Li W, Han J, Pei J. CMAR: Accurate and efficient classification based on multiple class－association rules[C]//ICDM'01. San Jose, CA: [s. n.], 2001: 369－376.

[10] Dong G, Zhang X, Wong L, et al. CAEP: Classification by Aggregating emerging patterns[C]//Proc. of the 2nd Int'l Conf. on Discovery Science (DS'99). Tokyo, Japan: [s. n.], 1999: 30－42.

[11] Fan H, Ramamohanarao K. A Bayesian Approach to use Emerging Patterns for Classification[C]//Proc of 14th Australasian Database Conference. Australia: Australian Computer Society, Inc, 2003: 39－48

[12] Dong G, Li J. Efficient mining of emerging patterns: Discovering trends and differences[C]//Proc. of KDD'99. San Diego, USA: [s. n.], 1999: 15－18.

(上接第 131 页)

1980.

[2] 韩东海，王　超，李　群. 入侵检测系统实例剖析[M]. 北京：清华大学出版社，2004.

[3] 邵峰晶，于忠清. 数据挖掘原理与算法[M]. 北京：中国水利水电出版社，2003.

[4] 熊忠阳，周亚峰. Web 访问挖掘的预处理技术的研究[J]. 计算机技术与发展，2007，17(8)：11－14.

[5] 孙吉贵，刘　洁，赵连宇. 聚类算法研究[J]. 软件学报，2008，19(1)：48－61.

[6] 韩家炜，堪　博. 数据挖掘：概念与技术. [M]. 第 2 版. 范明，孟小峰，译. 北京：机械工业出版社，2007.

[7] Li Lingjuan, Tang Wenyu, Wang Ruchuan. A CBR Engine Adapting to IDS[J]. Lecture Notes on Artificial Intelligence, 2005, 3802: 334－339, .

[8] MIT. MIT's KDD Cup 99 dataset[EB/OL]. 1999－10. http://kdd.ics.uci.edu/databases/kddcup99.html.

[9] 李玲娟，梁玉龙，王汝传. 适用于 IDS 中数据分类的数值归约算法[]. 计算机应用研究，2007，24(12)：146－148.

[10] 周桂芳. IDS 中特征选择算法的研究[D]. 南京：南京邮电大学，2006.

[11] 刘汉良. 统计学教程[M]. 第 3 版. 上海：上海财经大学出版社，2005.

[12] 梁玉龙. 数据挖掘技术及其应用研究[D]. 南京：南京邮电大学，2008.