基于商空间粒度聚类的异常入侵检测 *

王丽芳, 韩 燮
(中北大学 电子与计算机科学技术学院, 太原 030051)

摘 要: 针对异常入侵检测技术中传统聚类方法需要被检测类大小均衡的问题,在商空间粒度理论的基础上,论述了商空间粒度变换可以使复杂问题在不同的粒度世界求解,最终使整个问题得到简化。分析了商空间划分与聚类操作的相似性,提出了基于商空间的粒度聚类方法,并将该方法与入侵检测技术相结合,构建了基于商空间粒度聚类的入侵检测系统,用于对 KDD CUP 1999 数据集的异常入侵检测。实验结果表明,该入侵检测系统的性能明显优于基于传统聚类方法的入侵检测系统,从而证明了该方法的正确性和有效性。

关键词: 商空间; 粒度计算; 聚类; 异常入侵检测
中图分类号: TP393.08 **文献标志码**: A **文章编号**: 1001-3695(2010)05-1911-03
doi:10.3969/j.issn.1001-3695.2010.05.088

Anomaly intrusion detection based on quotient space granularity clustering

WANG Li-fang, HAN Xie
(*School of Electronics & Computer Science Technology, North University of China, Taiyuan* 030051, *China*)

Abstract: In view of the problem which traditional clustering methods need equilibrium detection class, this paper discussesd that quotient space granularity transformation could make complex problem to be solved in different granularity world basing on quotient space granularity theory, and ultimately simplified the whole problem. Then analyzed the similarity of the quotient space division and the clustering operation and put forward the method of granularity clustering based on quotient space. Moreover, combining the method and intrusion detection technology, constructed the intrusion detection system on the basis of quotient space granularity clustering and used the system to realize anomaly detection on the KDD CUP 1999 data sets. Finally experimental results show that the intrusion detection system is superior to other systems which is based on the traditional clustering methods. All these prove the correctness and effectiveness of the method.

Key words: quotient space; granularity computing; clustering; anomaly intrusion detection

0 引言

随着计算机技术和网络技术的不断发展,入侵技术越来越多样化,入侵检测系统所面临的数据也日益庞大。由于入侵检测系统中的异常检测本质上属于模式识别的范畴,而聚类分析是无监督模式识别的一个重要分支,将聚类算法应用于入侵检测系统得到了广泛的关注与研究[1,2],如美国哥伦比亚大学的 Portnoy 等人[1]较早提出了利用基于距离的聚类算法进行入侵检测,国内武汉大学罗敏博士实现了基于无监督聚类和支持向量机的网络入侵检测[3]。华中科技大学的李庆华教授等提出了基于聚类的有指导的异常检测方法[2]。本文针对传统聚类算法 K-means、FCM 算法存在的问题:只能检测类间大小近似相等的数据集,提出了一种基于商空间粒度聚类算法,并将该方法与入侵检测技术相结合,应用于入侵检测系统中。通过实验结果表明,该方法比现有的聚类算法具有更高的检测正确率。

1 相关知识

所谓粒度是一个物理概念,意指微粒大小的平均度量。张钹等人[4]提出了信息粒度的概念,是对信息和知识细化的不同层次的度量。信息粒度就是指人类在解决、处理和存储信息的有限能力上的一种反应,即人类在解决和处理大量复杂信息问题时,由于人类的能力有限,需把大量复杂信息按其各自特征和性能划分成数个较简单的信息块,以便处理,每个如此划分的信息块就被认为是一个粒度。在实际问题求解中,随着求解问题的不同,需要不同粒度的描述。

1.1 粒度的形式化描述

文献[5]使用一个三元组(X,f,T)描述一个问题。其中 X 表示问题的论域;$f(\cdot)$表示论域的属性,可用函数 $f:X\to Y$ 表示;T 是论域的结构,是指论域 X 中各元素的相互关系。分析或求解问题 (X,f,T),是指对论域 X 及其有关的结构、属性进行分析、研究。设 R 是 X 上的一个等价关系,则对 R 可以得到对应的商集,记为$[X]$。现在,在$[X]$上定义由 T 诱导出的拓扑,记为$[T]$,称$[T]$为商拓扑,$([X],[T])$为商拓扑空间。由拓扑学的原理知,从商空间的结构就可以了解原拓扑空间的某些性质。

从一个较粗的角度看问题,实际上就是对 X 进行简化,即把性质相近的元素看成是等价的,不但可把它们归入一类,并且可将整体作为一个新元素,这样就形成一个粒度较大的论域

收稿日期: 2009-08-27; **修回日期**: 2009-11-25 **基金项目**: 山西省自然科学基金资助项目(2007011042);中北大学青年科学基金资助项目(2008)

作者简介:王丽芳(1977-),女,山西长治人,讲师,博士研究生,主要研究方向为网络安全、粒度计算(wsm2004@nuc.edu.cn);韩燮(1964-),女,山西文水人,教授,博士,主要研究方向为数据挖掘、粒度计算.

$[X]$,也就把原问题转换成粗粒度上的问题($[X]$,$[f]$,$[T]$)。粒度与等价关系有着非常密切的联系,实际上,上面所说的简化过程和拓扑商集的概念完全相同。

1.2　不同粒度世界的关系

对同一个问题,有时需要在粗细不同的粒度世界中进行问题求解,则有必要建立不同粒度世界之间的关系。

定义 1　给定论域 X。设 R 是 X 上的一个等价关系,令 $[x]=\{y|yRx,y\in X\}$,$[X]_R=\{[x]|x\in X\}$,称 $[X]_R$ 为 X 对应于 R 的商空间,记为 $[X]_R$(或 $[X]$)。其中 xRy 表示 x 与 y 等价。

聚类操作与商空间的划分很相似。商空间的划分是在一等价关系上,将指定集合上等价的元素划分在一起,而聚类是将符合某种相似性的样本聚为一类,类内样本等价,类间样本相异。为此,聚类的粒度分析可从商空间不同粒度划分的角度来进行分析和讨论。

定义 2　设 R 表示由论域 X 上一切等价关系所组成的集合。可以如下定义等价关系,也就是粒度的粗和细,设 R_1,$R_2\in R$,如果对于任意元素 $x,y\in X$,都有 $xR_1y\rightarrow xR_2y$,那么就称 R_1 比 R_2 细,记为 $R_2<R_1$。

定理　R 在如上定义的"$<$"关系下可以形成一个完备半序格。

这个定理揭示了有关粒度的核心性质,因为其他性质都是以此为基础的,有关具体证明请参阅文献[5]。根据这个定理,可得到如下的序列:$R_0<R_1<\cdots<R_{n-1}<R_n$,直观地看,如上操作得到的序列和一棵 n 层树相对应。设 T(即论域的结构)是一棵 n 层的树,所有叶节点构成集合 X(论域),那么每一层节点都对应着 X 的一个划分。由于聚类操作得到的聚类谱系图恰好也是一棵 n 层树,必定存在一个等价关系序列与之对应,这也就是粒度和聚类之所以相通的原因。

2　粒度聚类分析法

引进粒度分析理论是为了有效地完成聚类任务,然而粒度取得太细,每个样本自成一类,不能挖掘样本中知识;粒度取得太粗,问题的某些性质被模糊。例如,学生的考试成绩如果是百分制,用 0~100 的整数来表示,如果是考查课,一般将其分成四个粒,分别用优、良、中、差来表示。优、良、中、差相对于百分制这个粒度空间来说它们是较粗的,而相对于四个粒的这个粒度空间来说,它们又是清晰的。选择合适粒度是聚类的关键。为了寻求合适的粒度,需要考察两个有益的等价划分。

定义 3　设 R_1 和 R_2 是论域 X 上的两个等价关系。如果 R 也是 X 上的一个等价关系,并且同时满足:$R_1<R$ 且 $R_2<R$;若有 R',使得 $R_1<R'$,$R_2<R'$ 且 $R<R'$;则称 R 为 R_1 和 R_2 之积,记为 $R=R_1\odot R_2$。

定义 4　设 R_1 和 R_2 是论域 X 上的两个等价关系。如果 R 也是 X 上的一个等价关系,并且同时满足条件:$R<R_1$,且 $R<R_2$;若有 R',使得 $R'<R_1$,$R'<R_2$,且 $R'<R$;则称 R 为 R_1 和 R_2 之和,记为 $R=R_1\ 8\ R_2$。

总之,$R_1\odot R_2$ 是能细分 R_1 和 R_2 最粗的划分,$R_1\ 8\ R_2$ 是能被 R_1 和 R_2 细分的最细的划分,即 $R_1\odot R_2$ 是划分 R_1 和 R_2 的最粗的上界,$R_1\ 8\ R_2$ 是划分 R_1 和 R_2 的最细的下界。

对一个具体问题聚类分析时,可参考图 1 所示方法。首先,根据问题需要预置一个等价关系 R_0 划分问题对应的集合(相应的粒度为 Q_0),得到商空间 S_0,在 S_0 上分析问题,得出初步结论 P_0 如果满足需要,聚类粒度合适,问题解决;否则,分以下两种情况考虑:

a)若与 Q_0 比较粒度偏粗,这时以 P_0 为指导,取一偏细等价关系 R_0',令 $R_1=R_0\odot R_0'$,在 R_1 上再进行分析,得出结论 P_1 和聚类粒度 Q_1,如果 P_1 还是粗的话,以 P_1 为指导,再取一偏细等价关系 R_1',令 $R_2=R_1\odot R_1'$,在 R_2 上再进行分析。以上过程可以重复进行,每重复一次,粒度将细化一次。

b)粒度偏细,取一偏粗等价关系 R_0',令 $R_1=R_0\ 8\ R_0'$,再进行分析,若还是细,再取偏粗等价关系 R_1',令 $R_2=R_1\ 8\ R_1'$,再进行分析。重复以上过程,可以逐渐加粗分析粒度。

利用上述粒度粗化和细化的手段,选择合适的粒度进行聚类的过程就是粒度聚类。

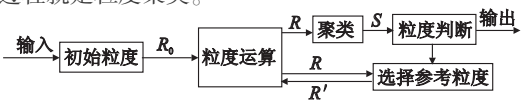

图1　粒度聚类过程

3　基于商空间粒度聚类的异常入侵检测方法

基于商空间粒度聚类的异常入侵检测系统主要由数据预处理、粒度聚类和检测系统三部分组成。系统组成结构如图 2 所示。

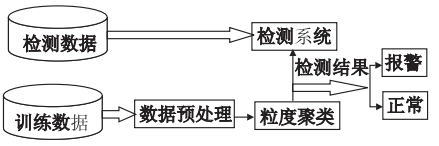

图2　基于商空间粒度聚类的异常入侵检测

数据预处理是在运用聚类算法之前,将原始数据转换成适合数据挖掘使用的格式,过滤和去掉噪声,即对数据的属性值进行标准化,为后期的粒度聚类做准备。粒度聚类处理模块对收集到的已经标准化的数据进行分类,区分哪些是正常的数据,哪些是异常数据,粒度聚类的结果将产生若干个簇,每个簇中包含部分的数据,正常的数据与异常的数据的特性不同,因此它们之间不具有相似性,应该处于不同的簇之中,这样就可以把包含异常数据的簇标记为异常簇,而将包含正常数据的簇标记为正常簇。完成对聚类结果的标志后,就可以实时地利用检测系统检测数据了,根据检测算法进行增量聚类,在不断完善聚类结果的同时,有效地检测出未知入侵行为。整个检测过程分为两个阶段,即训练阶段和检测阶段。在训练阶段,粒度聚类对网络数据进行处理,得出聚类中心;在检测阶段,检测系统根据粒度聚类提供的聚类中心和设定的阈值对测试数据进行判断。

算法　基于商空间粒度聚类的异常入侵检测算法。

输入:网络连接数据集 X,数据的属性集 A。

输出:检测结果。

a)对训练数据 X 进行预处理[6],删除重复和多余的属性 a,即 $A=A-a$。

b)计算数据集超球体的半径 R,作为初始粒度(即也是最粗粒度),并置聚类中心代表点集 M 为空。

c)以 R 为参数进行聚类,并将结果添加到 M 中。

d)以 M 为聚类中心生成聚类簇,以给定的相似度函数 $f=1/E$。其中 E 的计算如文献[7],及相似度阈值 s。考察每个簇中数据的分布情况:如果聚类簇中数据的相似度 $f>s$,则将该

聚类中心保留在 M 中;否则将聚类簇中的数据添加到更小粒度聚类的数据集 X_1 中。

e)如果数据集 X_1 为空集,则聚类结束,否则令 $R = R \times \lambda$,其中 λ 为聚类下降系数,其取值大小如文献[8],转至 c)。

f)将聚类得到的 M,加入到检测系统的规则库 J 中,即 $J = J + M$,其中 K 的初值为 0。

g)检测系统依据其内部的规则,对检测数据进行检测。检测方法为:采用欧氏距离 $d(i,k)$ 计算值的倒数来表示被检测数据与规则库 K 中规则的接近程度,假如规则库中有三个规则,则分别求出 K_1、K_2、K_3,其中 $K_i = 1/d(i,k)$,$(i = 1,2,3,\cdots)$。K_i 最大的表示待测的数据最靠近第 i 个规则,如果 i 规则为正常模式,则待测数据为正常数据;反之则为异常数据,系统发出报警。

h)算法结束。

4　实验结果

在实验过程中,由于实验室的条件有限,无法收集到网络中各种攻击数据,对数据的训练是基于入侵检测经典数据集 KDD CUP 1999[9]。KDD CUP 1999 数据集中的入侵主要分为四类:拒绝服务类型 DoS,其攻击目的是破坏或拒绝合法用户对网络、服务器、服务或其他资源的访问;远端机器未授权登录访问 R2L;未经授权且试图获取超级用户和 root 权限类型 U2R;对弱点的监视或其他探测类型 Probe。由于原始数据集大约有 500 万条记录,过于庞大,在实际实验中选取 KDD 中的两子集来测试算法的性能,这个子集包含了全部的攻击类型,入侵数据约占总数的 2% ~2.5%。具体实验数据如表 1 所示。

表1　实验数据

攻击类型	数据集1	数据集2
记录个数	11 000	12 000
攻击总次数	250	250
DoS攻击次数	180	100
R2L攻击次数	35	50
U2R攻击次数	5	10
Probe攻击次数	30	50
未知攻击（如Satan、Neptume等）	0	40

实验中把数据集 1 作为训练集,数据集 2 作为检测集。为了验证该方法对未知攻击的检测能力,在数据集 2 中包含了数据 1 中没有出现的攻击类型,如 Satan,Neptune 等。入侵检测的过程是首先用聚类算法在数据集 1 上进行训练,得到各聚类的结果,从而构建了正常与异常行为的规则库,通过此规则库就可以对检测数据进行入侵检测了。入侵检测系统的性能主要由检测率 DR 与误检率 FR 两个方面的值来体现。其中,DR = 检测到的入侵样本数/入侵样本总数;FR = 被误报为入侵的正常样本数/正常样本总数。实验中使用 K-means、FCM、粒度聚类三种算法分别对数据集 1 和 2 进行了聚类和检测,并对三种算法在聚类准确率、训练结果和检测结果、各攻击类型(DoS、Probe、U2R、R2L)检测率等方面进行了比较,比较结果如图 3 ~6 所示。其中图中结果取自 10 次实验的平均值。从实验结果可以看出基于商空间的粒度聚类方法在聚类准确率、对已知攻击的检测率及对未知攻击的检测率等方面都明显优于传统的 K-means、FCM 算法,究其原因在于异常网络入侵数据集中,正常数据所占的比例远远大于异常数据的比例,而传统的 K-means、FCM 算法只能进行各个类大小均衡的数据集的聚类,所以将这些算法应用于入侵检测中得到的实验结果较差。

5　结束语

本文针对入侵检测中所使用的传统聚类方法存在检测结果不准确的问题,提出了将基于商空间的粒度聚类方法与入侵检测技术相结合的基于商空间粒度聚类的异常入侵检测方法,并应用此方法对经典的网络入侵数据集 KDD CUP 1999 进行了实验,将实验结果与传统的聚类方法 K-means,FCM 的实验结果进行了比较,表明该方法在聚类准确性及检测效率等方面均明显优于传统的聚类方法,说明该方法具有一定的可行性。

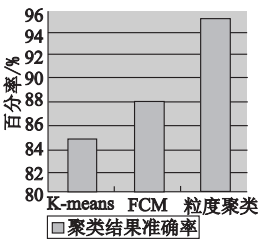

图3　三种算法聚类准确率比较

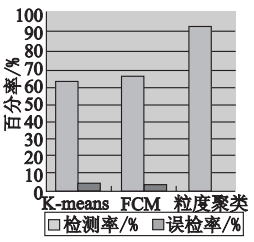

图4　三种算法对数据集1的训练聚类结果比较

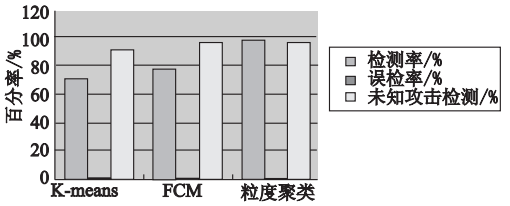

图5　三种算法对数据集2检测聚类结果比较

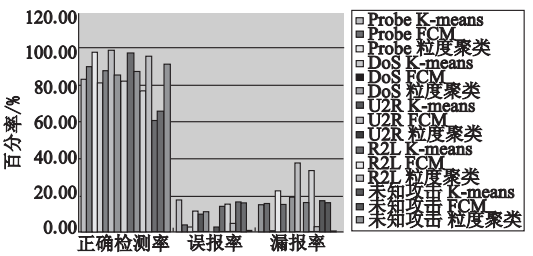

图6　各攻击类型检测率比较

参考文献:

[1] PORTNOY L, ESK N E, STOLFO S J. Intrusion detection with unlabeled data using clustering[C]// Proc of ACM CSS Workshop on Data Mining Applied to Security. New York:ACM,2001.

[2] CH M PHLEE W, ABDULLAH A H, NOOR M S M, *et al*. Integrating genetic algorithms and fuzzy C-means for anomaly detection [C]//Proc of Annual IEEE NDICON. Washington, DC: IEEE, 2005:575-576.

[3] KRISHNAPURAM R, KELLER J M. A possibilistic approach to clustering[J]. IEEE Trans on Fuzzy System,2003,1(2):87-88.

[4] 张钹,张铃.问题求解理论及应用[M].北京:清华大学出版社,1990:45-67.

[5] 王珏,苗夺谦,等.关于 Rough Set 理论与应用的综述[J].模式识别与人工智能,2006,9(4):337-344.

[6] LAZAREVIC A, ERTOZ L, KUMAR V, *et al*. Data mining:a comparative study of anomaly detection schemes in network intrusion detection[C]// Proc of the 3rd SIAM International Conference. Rotterdam:[s.n], 2003.

[7] 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,2003:10-50.

[8] STALLINGS W. High-speed networks and Internets: performance and quality of service[M]. 2nd ed. New York: Prentice-Hall, 2002: 148-152.

[9] The third international knowledge discovery and data mining tools competition dataset [DB/OL]. (1999-10-28)[2009-03-05]. http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html.