基于K-MEANS聚类的分支定界算法在网络异常检测中的应用

杨宇舟[1] 张凤荔[2] 王 勇[2]
(电子科技大学软件学院 成都 611731)[1] (电子科技大学计算机科学与工程学院 成都 611731)[2]

摘 要 网络异常检测技术是入侵检测领域研究的热点之一。在异常检测中,针对其存在的对训练集中关键数据的选取不准确、选取过程耗时较长、检测的误报率过高等问题,结合经典的K-MEANS算法和分支定界算法,建立起一种网络异常检测模型,以有效地提高在大量训练集中选取关键数据的准确率,同时降低数据选取的时耗。通过大量基于著名的KDD Cup 1999数据集的实验,表明此模型能够达到较高的检测准确性,并能有效地控制检测错误报警的发生。

关键词 异常检测,K-MEANS,分支定界
中图法分类号 TP309.2 **文献标识码** A

Application of Branch and Bound Algorithm Based on K-MEANS Clustering in Network Anomaly Detection

YANG Yu-zhou[1] ZHANG Feng-li[2] WANG Yong[2]
(School of Software, University of Electronic Science and Technology of China, Chengdu 611731, China)[1]
(School of Computer Science & Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China)[2]

Abstract Network anomaly detection has become one of the focus research topics in the field of intrusion detection. However, issues on accurate selection of key date in training set, the long selection time, and the high rate of detection misstatement are still unresolved. Regarding to those problems, to integrate K-MEANS and Branch and Bound Algorithm, and to build up a network anomaly detection model on it can significantly improve the accuracy of key data selection, and reduce time consumption as well. A series of experiments on well known KDD Cup 1999 dataset demonstrate that the model can achieve a high detection accuracy and efficiently constrain the false alarms caused by detection.

Keywords Anomaly detection, K-MEANS, Branch and bound

1 引言

异常检测(Anomaly Detection)[1]是假设入侵者的行为在正常用户的行为范围之外,它需要建立系统正常活动状态或者正常用户行为模式的描述模型,然后将当前的用户行为模式或者当前的系统状态与该正常模型进行对比,看当前值是否超出了预设的阈值。如果超出,则认为存在入侵行为[2]。异常检测的难点在于如何建立正常用户的行为模式或者正常系统状态的描述模型以及设计统计算法,避免把正常行为误作入侵或者入侵行为误作正常[3]。在大量的网络数据中,要建立正常用户的行为模式需要对数据进行有效筛选,因此选取合理的数据挖掘[4]、聚类算法便显得尤为重要。

聚类分析(cluster analysis)是一种无监督的学习方法,它是将物理或抽象对象的集合分组成由类似对象组成的多个类的过程,其目标是让同一类中的数据具有较高的相似性。聚类分析可以发现属性之间存在的联系,从而找出数据分布的模式[6]。

在数据的异常检测中,搜索某点的K近邻点并进行检测分析,是一个耗时的过程。K-MEANS算法[7]是数据挖掘的经典算法[8]。分支定界算法是一种有效的查找算法,将分支定界算法运用在K-MEANS算法的基础上,可以大幅提高查找的效率,减少不必要的消耗;将其应用于异常检测技术中,能够及时、高效、精确地识别异常行为,使整个检测保持较高的准确率,同时将误报率控制在较低的范围内。

2 基于K-MEANS的分支定界算法

网络上的数据量是复杂的、巨大的,如何从庞大的数据集中快速、准确地找出异常检测活动训练集中最有代表性的数据来进行分析、比较、判断,从另一方面决定着一个异常检测方法的检测效率和检测准确度。对异常检测算法的优化不仅体现在异常检测本身的算法上,同时体现在对异常检测所需数据采集的数据挖掘[5]算法上。K-MEANS算法是数据挖掘中的层次聚类算法,它将数据分成多个层次,然后对不同层次的数据采用划分聚类,输出一棵层次化的分类树。将所有对象置于一个簇中,自顶向下逐渐细分为越来越小的簇,直至达到某个终结条件。分支定界算法对生成的分类树进行搜索查

到稿日期:2011-05-20 返修日期:2011-08-12 本文受信产部电子发展基金(信部运(2007)329),四川省科技厅基金(M110106012009FZ0148),国家科技部科技人员服务企业行动项目(SQ2009GJE0000163)资助。

杨宇舟(1986—),男,硕士生,主要研究方向为计算机网络安全,E-mail:gecko_21@163.com;张凤荔(1963—),女,教授,博士生导师,主要研究方向为网络安全、移动数据管理及其应用等;王 勇(1976—),男,副教授,硕士生导师,主要研究方向为P2P网络安全。

找，找到距离目标点最近的 K 邻近点。两种算法结合起来，可以高效、准确地提取活动训练集中的代表性数据，再将其应用到异常检测中，使整个异常检测具有较高的准确率。

2.1 K-MEANS 分裂算法构建数据树形结构

K-MEANS算法可以快速地将大量数据分类聚合，使数据的查找更加准确、快速。算法以 k 为输入参数，把 n 个对象集合分成 k 个簇，使得结果簇类内部的相似度高，而簇间的相似度低[9]。

1)一些参数的定义

S_P：和结点 P 有关的样本集合。

N_P：和结点 P 有关的样本数。

M_P：S_P 的样本均值。

r_P：从 M_P 到 $X_i \in S_P$ 的最远距离：$r_P = \max d(X_i, M_P)$ $(X_i \in S_P)$。

2)算法描述

输入：簇的数目 k 和包含 n 个对象的数据集以及划分的树形结构的层数 L。

输出：具有 n^L 个叶子结点的树形结构。

步骤：

①选择 k 个对象作为初始的簇中心；

②计算 n－k 个对象与这 k 个对象的距离，按最短距离将对象划分到 k 个簇类中，并计算每一簇类的 N_P、M_P 和 r_P；

③重复步骤②，将第 1 层的 k 个簇类各自分别再划分为 k 个类，作为树形结构的第 2 层；不断划分，直至划分为所需要的 L 层，使第 L 层有 k^L 个叶子簇类时算法结束。

本节为方便说明，将 k 取值为 3，L 取值为 3，基于 K-MEANS 的分裂算法生成的树形结构如图 1 所示。

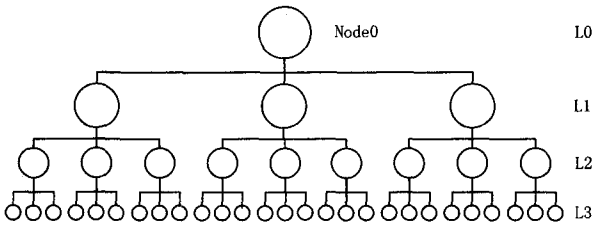

图 1　以 3 个对象作为簇中心的 3 层树形结构

此阶段经过多次迭代建立起了数据的树形结构，将数据分为多个聚类，同一类中的数据具有较高的相似性。

2.2 分支定界算法搜索 K-近邻

上一节将大量数据分为了 N 个类，并计算出了 N_P、M_P 和 r_P。接下来需要搜寻与已知点相似度最高的 K 个点，即 K-近邻。寻找 K-近邻是一个相当耗时的过程，将分支定界的方法运用于查找 K-近邻的过程中，能高效、快速地搜寻出目标对象。分支定界算法的原理如下[10]：

1)两条规则的说明

规则 1　如果

$$B + r_P < d(X, M_P) \tag{1}$$

成立，则不存在 $X_i \in S_P$ 使 X_i 到 X 的距离是最短距离（B 是样本中到 X 的当前最短距离。最开始时，令 B 的值为∞）。

规则 1 的证明如图 2 所示。

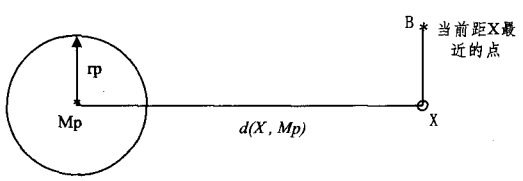

图 2　规则 1 的说明图

对于任意的 $X_i \in S_P$，由几何定理可知，$d(X, X_i) + d(X_i, M_P) \geqslant d(X, M_P)$；

因此由定义，$d(X_i, M_P) \leqslant r_P$，得

$d(X, X_i) \geqslant d(X, M_P) - r_P$；

所以，如果 $d(X, X_i) \geqslant d(X, M_P) - r_P > B$ 成立，则不存在 $X_i \in S_P$ 使 X_i 到 X 的距离是最短距离。

规则 2　如果

$$B + d(X_i, M_P) < d(X, M_P) \tag{2}$$

成立，则 X_i 到 X 的距离不是最短距离。$X_i \in S_P$。

规则 2 的证明如图 3 所示。

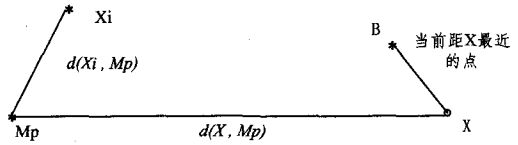

图 3　规则 2 的说明图

2)分支定界算法搜索 K-近邻的步骤

以上一节建立的树形结构为基础。

①(初始化)：设置 $B = \infty$，CURRENT LEVEL $L = 1$，CURRENT NODE=0；

②(分裂当前结点)：把 CURRENT NODE 分裂出的所有结点添加到 CURRENT LEVEL L 对应的 ACTIVE LIST 链表中，并计算这些点的 $d(X, M_P)$；

③(用规则 1 进行检测)：检测 CURRENT LEVEL L 对应 ACTIVE LIST 链表上的每一个 P 结点，如果 $B + r_P < d(X, M_P)$ 成立，则从 ACTIVE LIST 链表上删除该 P 结点；

④(回溯)：如果 CURRENT LEVEL L 对应的 ACTIVE LIST 链表上的结点数为 0，返回上一级的 LEVEL，设置 $L := L-1$。如果 $L=0$，算法到此结束；如果 $L \neq 0$，返回步骤③继续进行；如果 CURRENT LEVEL L 对应 ACTIVE LIST 链表上的结点数为 0，则执行步骤⑤；

⑤(选择距离最短的结点进行分裂)：在 CURRENT LEVEL L 对应 ACTIVE LIST 链表上，选择 $d(X, M_P)$ 最小的结点 P 设为当前结点 CURRENT NODE；从 CURRENT LEVEL L 对应 ACTIVE LIST 链表上删除结点 P。如果 CURRENT LEVEL L 是最后一层（L 为最大值），则执行步骤⑥；否则，设置 $L := L+1$，执行步骤②；

⑥(用规则 2 进行检测)：对 CURRENT NODE P 中的每一个元素 X_i，执行以下的检测：如果 $d(X, M_P) > B + d(X_i, M_P)$，说明 X_i 不是距 X 最近的元素，不用计算 $d(X, X_i)$；否则计算 $d(X, X_i)$。如果 $d(X, X_i) < B$，设置 CURRENT NN=i，$B = d(X, X_i)$。在所有的 X_i 都检测完后，回到步骤③执行。每得到一个新的 B，就与 CURRENT K-NEAREST NEIGHBOR TABLE 中的数据比较，如果有大于 B 的元素存在，则将 B 添加到 CURRENT K-NEAREST NEIGHBOR TABLE 中，并删除 CURRENT K-NEAREST NEIGHBOR TABLE 中的最大值。

算法结束后，CURRENT K-NEAREST NEIGHBOR TABLE 中的元素便是要求得到的 K-临近元素（K-Nearest Neighbors）。

3　基于分支定界算法的网络异常检测模型

本节根据分支定界算法，结合直推式异常检测方法[11]，

构建了一种网络异常检测模型，旨在说明如何在实践中使用此方法进行异常检测。模型如图 4 所示。

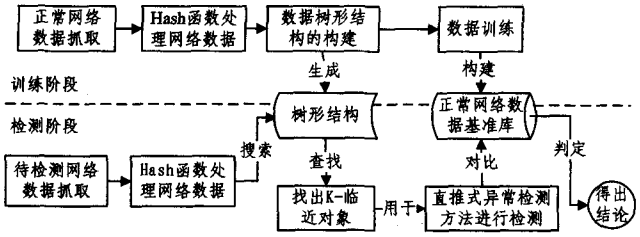

图 4　基于分支定界算法的网络异常检测模型

在图 4 的上半部分，训练阶段为了建立实际应用中的正常网络训练集，所要进行的操作主要包括以下几部分：

· 抓取正常网络数据：从网络中采集、选取能够反映正常行为的数据，用以构建检测阶段用于异常检测的正常行为数据集；

· Hash 函数处理网络数据：将网络的各种特征数据转换为可操作的数据用于算法分析；

· 构建数据树形结构：运用 K-MEANS 聚类算法将网络数据分类，构建出树形结构，为数据训练做好准备；

· 训练数据：就之前准备好的网络数据进行训练，构建出正常数据基准库。

在图 4 的下半部分，给出了检测阶段的主要工作：

· 抓取待检测网络数据：主要负责从监控的目标网络中收集原始网络数据；

· Hash 函数处理网络数据：将网络的各种特征数据转换为可操作的数据，用于算法分析；

· 找出 K-邻近对象：运用分支定界算法，利用训练阶段构建出的树形结构，查找出待检测数据的 K-Nearest Neighbors，用于异常检测算法；

· 直推式异常检测方法进行检测：将数据代入直推式异常检测算法[11]，对比正常网络数据基准库，计算得到 P 值，判定出此检测数据是否异常。

4　实验及其结果分析

本节将对所提出的异常检测模型的可行性和有效性进行验证。为了保证实验的说服力及方便性，本节采用研究领域共同认可并广泛应用的基准评测数据集 KDD Cup 1999 进行测试。本文采用的评价指标为国际上通用的检测率（true positive rate，简称 TP）和误报率（false positive rate，FP）。

4.1　实验数据集

本文采用的 KDD Cup 1999 数据集包括大约 4900 000 条数据记录，每条都是从军方网络环境中模拟攻击所得的原始网络数据根据设定的 41 个特征提取出来的，它们都是描述网络连接统计信息的特征向量，包含 5 类数据：DoS，Probe，R2L，U2R 这 4 类攻击数据（共包含 24 种攻击类型）以及正常数据[11]。

4.2　实验结果及分析

我们将 KDD Cup 1999 的数据集进行了提取。本文随机从数据集中提取了 7164 条正常数据，用于建立标准训练集，又随机提取了 834 条正常数据和 554 条攻击数据（包含以上 4 类攻击）进行检测。

实验中，本文将异常检测的置信度设为 0.05，用正常数据和攻击数据从正反两个方面验证了检测方法、模型的有效性和检测准确性。实验结果如表 1 所列。

表 1　正常数据和攻击数据的检测率

K Neighbor Value Set	Data Type	Number of Attack Alarm	FP(%)	TP(%)
K=3	Attack	548/554	1.0830	98.9170
	Normal	61/834	7.3141	92.6859
K=5	Attack	547/554	1.2635	98.7365
	Normal	48/834	5.7554	94.2446
K=8	Attack	546/554	1.4440	98.5560
	Normal	36/834	4.3165	95.6835
K=13	Attack	544/554	1.8055	98.1945
	Normal	35/834	4.1966	95.8034

表 1 中 Attack 表示攻击类型的数据包，用来试验算法的检测准确性；Normal 表示正常的数据包，用来试验算法的检测误报率。

K 的值代表算法中选取的邻近结点的数目，这些结点是距离某已知结点最近的 K 个点。

为了试验 K 的选取对算法的影响，本文选取了 4 个不同的 K 值进行实验检测。从表 1 的实验结果中可以看出，当 K 值较小时，对攻击数据的检测较为准确，但对正常数据检测的准确率则相对较低，因为此时的 K 个结点距已知结点的平均距离较近，算法对于数据较为敏感，容易检测出攻击数据，也容易把正常数据误当作异常数据；当 K 值增大时，对攻击数据的检测准确率略有下降，但对正常数据的检测准确率则稍微上升，因为此时的 K 个结点距已知结点的平均距离不那么近，算法对于数据的敏感度也不那么高，不容易把正常数据包误作为异常包而告警，也没那么容易检测出攻击、异常包。但这样的差别只是微小的，不会影响到检测的功能。K 值越大，检测消耗的时间越多。在实际应用中，应该综合各方面因素，选择一个较为合适的 K 值，以达到最理想的检测效果。

本文所述方法的最大优点是：不需要对攻击方式进行建模和学习（实践中也很难比较全面地获得这些攻击数据），只需要对“相对干净”的正常数据进行学习和检测，因此在实践中更为实用。本文的实验结果也证明了其高效性和可行性。

结束语　本文结合数据挖掘的特点和优势，将 K-MEANS 算法和分支定界算法结合起来，运用到网络异常检测领域当中，提出了一种异常检测方法以及模型。在经典的 KDD Cup 1999 数据集上的实验表明，本方法具有较高的检测准确率和较低的误报率。

本方法在实践应用中还需根据实际情况做进一步改进，以提高性能。在数据选取方面，应考虑如何有效地从网络数据中选取一定量的具有代表性的精简特征，避免由冗余导致算法效率下降；在训练集的建立过程中，应考虑如何提取少量的关键数据，以降低算法运算的时间复杂度，减少时耗等。

参 考 文 献

[1] Grzech A. Anomaly detection in distributed computer communications systems[J]. Cybernetics and Systems, 2006, 37(6): 635-652

[2] Chandola V, Banerjee A, Kumar V. A Survey on Anomaly Detection[R]. University of Minnesota, 2008

[3] Fu De-sheng, Zhou Shu, Guo Ping. Design and Implementation of Distributed Network Intrusion Detection System Based on Data Mining[J]. Computer Science, 2009, 36(3): 103-105

（下转第 97 页）

上述典型方案和本文方案均实现了数字内容及其相应许可的安全分发，前者主要采用传统的数字内容加密方法，而本文则基于 AP^2RA 远程证明方法，结合可信计算用户终端平台实现多媒体内容的安全、可信分发；在系统开销上，本文方案采用了第三方可信平台验证机制，因此开销略大，但更适合于一般开放网络。

结束语 为解决多媒体社交网络应用下的数字内容版权管理问题，本文将可信计算中的远程证明技术引入到 DRM 安全方案中，提出了一种基于支持验证方代理的数字内容分发方案与安全协议。通过与其他代表性的 DRM 方案相比，本文方案提高了数字内容分享和传播的安全性、可信性和可控性，从而满足了 MSN 中用户节点终端平台的隐私保护这一实际需求。进一步工作将对本文协议进行形式化验证分析，并将其应用于普适的社交网络信息交换与共享应用服务中。

参 考 文 献

[1] Rosenblum D. What anyone can know: The privacy risks of social networking sites [J]. IEEE Security and Privacy, 2007, 5 (3): 40-49

[2] 张志勇，牛丹梅. 数字版权管理中数字权利使用控制研究进展[J]. 计算机科学，2011，38(4)：48-54

[3] 邱罡，王玉磊，周利华，等. 基于可信计算的 DRM 互操作研究[J]. 计算机科学，2009，36(1)：77-80

[4] Grawrock D. TCG Specification Architecture Overview Revision 1. 4 [EB/OL]. https://www. trustedcomputionggroup. org/groups/ TCG_1. 4_Archiecture_Overview. pdf, 2011-05-01

[5] 谭良，刘震，周明天. TCG 架构下的证明问题研究及进展[J]. 电子学报，2010(5)：1105-1112

[6] 张焕国，陈璐，张立强. 可信网络连接研究 [J]. 计算机学报，2010，33(4)：706-717

[7] 张志勇，裴庆祺，等. 支持验证代理方的远程证明模型及其安全协议[J]. 西安电子科技大学报，2009，36(1)：58-63，105

[8] Popescu B, Crisop B, Tanenbaum A, et al. A DRM security architecture for home networks [C] // Proceedings of 4th ACM Workshop on Digital Rights Management. Oct. 2004

[9] Kim H, Lee Y, Chung B, et al. Digital Rights Management with right delegation for home networks [C] // Proceedings of 9th International Conference on Information Security and Cryptology. LNCS 4296, 2006: 233-245

[10] 马兆丰，范科峰，陈铭，等. 支持时空约束的可信数字版权管理安全许可协议[J]. 通信学报，2008，29(10)：153-164

(上接第 62 页)

[4] Lee W, Stolfo S J. A Data mining framework for building intrusion detection models [M]. IEEE Computer Society Press, 1999: 120-132

[5] Yue Yao-xue. The Research of Network Intrusion System Based on Algorithm of Data Mining[J]. Comput-er Security, 2009, 10: 41-43

[6] Liu He-bing, Shang Jun-ping. On Clustering Analysis Algorithm [J]. Journal of Yiyuan Vocational and Technical Collage, 2006, 5 (4): 4-7

[7] Hartigan J A, Wong M A. A K-Means Clustering Algorithm [J]. Journal of the Royal Statistical Society, Series C (Applied Statistics), 1979, 28(1): 100-108

[8] Xindong W, Kumar V, Quinlan J R, et al. Top 10 Algorithms in Data Mining[J]. Knowledge and Information Systems, 2008, 14 (1): 1-37

[9] Li Ling-juan, Li Bing, Xue Ming. Research on Application of K-MEANS Algorithm in IDS[J]. Computer Technology and Development, 2010, 20(7): 129-131

[10] Fukunaga K, Narendra P M. A Branch and Bound Algorithm for Computing K-Nearest Neighbors [J]. IEEE Transactions on Computers, 1975(7): 750-753

[11] Li Yang, Fang Bin-xing, Guo Li, et al. A Network Anomaly Detection Method Based on Transduction Scheme[J]. Journal of Software, 2007, 18(10): 2595-2604

(上接第 78 页)

结束语 本文主要针对物联网环境中的服务获取问题提出了一种基于人工能量势的空间社区服务获取方法。物联网中的节点在进行服务获取时，利用节点最大有效传输范围选择下一跳节点，提高了节点生存周期，也在一定程度上增强了服务获取的效率。但是由于空间社区中节点的多样性和海量性，导致在一定程度上影响到服务获取的安全性和发现效率，我们下一步工作将集中于如何构建物联网实验床并开发基于人工能量势的服务获取原型系统。

参 考 文 献

[1] Commission of the European Communities. Internet of Things-An Action Plan for Europe (1st Edition) [R]. Brussels: COM, 2009, 278: 1-12

[2] Yu Tao, Zhang Yue, Lin K-J. Efficient Algorithms for Web Services Selection with End-to-End QoS Constraints[J]. ACM Transactions on the Web, 2007, 1(1)

[3] 王杨，王汝传. 一种基于 Echord 协议的网格资源发现方法[J]. 电子学报，2010，38(11)：2499-2504

[4] Zhao Q, Liu J, Xu J. Improving Search on Gnutella-like P2P Systems [C] // Computational Science-ICCS2007-7th International Conference. Berlin Heidelberg: Springer-Verlag, 2007, 4490(4): 887-890

[5] Khatib O. Real-time obstacle avoidance for manipulators and mobile robots [J]. The International Journal of Robotics Research, 1986, 5(1): 90-98

[6] Zhong M, Shen K, Seiferas J. The convergence-guaranteed random walk and its applications in peer-to-peer networks [J]. IEEE Transactions on Computers, 2008, 57(5): 619-633