基于聚类融合的入侵检测

李 建,李 杰,孙燕花
(中南大学 信息科学与工程学院,湖南 长沙 410083)

摘 要:随着互联网的飞速发展,网络安全的问题日趋严重,传统的网络安全技术已难以应对日益繁多的网络攻击。因此入侵检测便应运而生了,而且其重要性日益提高。基于聚类分析的入侵检测已经成为其主要研究方向。聚类分析是一种有效的异常入侵检测方法,可用以在网络数据集中区分正常流量和异常流量。但单一的聚类算法很难达到预期的效果,为了提高入侵检测的效果,文中采用聚类融合技术,提出一种基于 Co-assocition 的模糊聚类融合算法,通过实验检测能显著提高检测率和降低误报率。

关键词:网络安全;入侵检测;聚类融合
中图分类号:TP393.08 **文献标识码**:A **文章编号**:1673-629X(2011)10-0250-04

An Intrusion Detection Based on Clustering Ensemble

LI Jian, LI Jie, SUN Yan-hua
(School of Imformation Science and Engineering, Central South University, Changsha 410083, China)

Abstract: With the rapid development of network, more and more network security problems are appearing, the traditional network security technology has been difficult to protect the network by growing range of network attacks. So the intrusion detection is turned out, and it gets more important in the network. Intrusion detection based on cluster analysis has become the main research directions. Cluster analysis is an effective method for anomaly intrusion detection, and it can distinguish the normal and abnormal data of the network data. But a single clustering algorithm is hard to achieve the desired effect. In order to improve the effectiveness of intrusion detection, proposes a new fuzzy clustering ensemble algorithm based on Co-assocition . Through experimental testing can significantly improve the detection rate and lower false alarm rate.

Key words: network security; intrusion detection; clustering ensemble

0 引 言

入侵检测[1]是一种主动的安全措施,它从系统内部或网络中收集信息,并对这些信息进行检测以便得出是否存在网络入侵。对于"入侵"的定义目前有许多提法,但较为公认的是美国国家安全通信委员会(NSTAC)的入侵检测小组(IDSG)于1997年给出的定义。入侵是对信息系统的非授权访问或者未经许可在信息系统中进行的操作。同年该组又给出了入侵检测的定义。入侵检测:是指对企图入侵、正在进行的入侵或者已经发生的入侵进行识别的过程。因此入侵检测系统是指:能够执行入侵检测任务和功能的系统。

入侵检测系统按照数据源的来源可分为基于主机的入侵检测系统和基于网络的入侵检测系统[2]。基于主机的入侵检测也称基于系统的模型,它是通过分析系统的审计数据来发现可疑的活动,如内存和文件的变化等。其输入数据主要来源于系统的审计日志,一般只能检测发生在该主机上的入侵。基于网络的入侵检测系统使用原始的网络数据包作为信息源。通过将主机的网卡设成混杂模式,实时监视并分析在网络中传输的数据包,如果发现了攻击行为,入侵检测系统的响应模块就会对攻击行为采取相应的措施。随着经济的高速发展,我国的网络技术有了长足的发展。与此同时,网络安全问题也日趋严重。传统的入侵检测技术已难以应对复杂多变的网络攻击,于是聚类分析[3]应用到了入侵检测系统中。

1 聚类概述

聚类是一项重要的发现数据分布和隐含模式的数据挖掘技术[4]。但是单一的聚类算法很难达到预期的效果[5]。故文中采用聚类融合技术,以提高算法的稳定性。

聚类也称无监督的分类,聚类就是将一个数据划分为多个类或簇的过程,并使得同一个类的数据对象具有较高的相似性,而不同类中的数据差别比较大。

收稿日期:2011-03-21;**修回日期**:2011-06-24
作者简介:李 建(1986-),男,湖南湘潭人,硕士研究生,研究方向为网络管理;李 杰,教授,研究方向为网络管理。

定义 1 聚类分析的输入可以用一组有序对 (X,s) 或 (X,d) 表示,其中用 X 表示一组样本,而 s 和 d 则用来分别表示样本间相似度或相异度(距离)。其聚类结果也是输出一个分区,若 $C=\{C_1,C_2,\cdots,C_k\}$,其中 $C_i(i=1,2,\cdots,k)$ 是样本集 X 的一个子集,可以用如下所示:

$$C_1 \cup C_2 \cup \cdots \cup C_k = X$$

$$C_i \cap C_j = \varnothing, i \neq j$$

其中 C 中的每个成员 $C_1,C_2,\cdots C_k$ 都叫做一个类,每一个类都是使用一些符合人们思维的特征进行描述的,一般由类的中心或类的边界点代表一个类;使用聚类树中的结点图形化地表示一个类的情况也很多;使用样本属性的逻辑表达式也较为常见。其中,用类的中心点表示一个类最为常见。

2　聚类融合综述

聚类融合是利用不同的算法或同一算法下的不同参数得到的聚类结果进行融合[6]。文中对国内外聚类融合现状进行了综述,并在入侵检测系统中采用基于 Co-association 的聚类融合算法,实验证明聚类融合算法比单一的聚类算法优越。聚类融合(Clustering Ensemble)最早是由 A. Strehl 和 J. Ghosh 在文献[7]中提出的。在文中,作者将聚类融合定义为:将多个对一组对象进行的不同结果进行合并,而不使用对象原用的特征[7]。

具体表达如下:假设有 n 个数据点的集合 $X=\{x_1,x_2,x_3,\cdots x_n\}$,对数据集 X 用 M 次聚类算法得到 M 个聚类结果 $M=\{Q_1,Q_2,Q_3,\cdots Q_m\}$,其中 Q_i(其中 $i=1,2,3\cdots m$)为运行第 i 次聚类算法得到的结果,然后通过设计一种共识函数 F,以期实现对这 M 个聚类结果的聚类融合,得到一个聚类结果 Q,如图 1 所示。

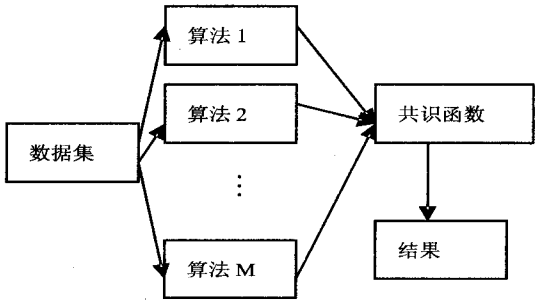

图 1　聚类融合示意图

聚类融合算法比单一聚类算法有明显优势,在文献[8]中,A. Topchy 总结出了如下几点:

(1)适用性。由于单一的聚类算法局限较大,故聚类融合的适用性更强。

(2)鲁棒性。由于聚类融合不容易受参数的影响太大,故在各种各样的领域和数据上的平均性能更为优越。

(3)并行性和可扩展性。能对数据子集进行并行聚类并进行合并;聚类融合比单一的聚类算法具有更好的并行性和可扩展性。

(4)稳定性。由于多次运行聚类算法故对噪声的处理较好,而且孤立点和抽样方法等对聚类结果的影较小。

聚类融合的过程是:首先对样本集进行多行聚类产生聚类成员,其次对聚类成员融合得出聚类结果。故当前聚类融合研究的主要方向是:

1. 怎样产生高效的聚类成员;

2. 怎样的共识函数才能产生高效的聚类融合算法。

B. Minaei-Bidgoli 在文献[9]中对聚类融合的研究方向作了详细说明。主要包括两大方向,聚类成员的产生和共识函数的设计。聚类成员产生可以通过采用不同算法、同一算法不同参数、不同数据字集、不同特征子集、投影到子空间等方法来产生。共识函数的设计可以采用 Co-association 方法、投票方法、信息论方法、超图方法[10]、混合模型等方法实现。

3　一个改进的基于 Co-association 的聚类融合算法(C-EFC)

基于 Co-association 的聚类融合算法,首先得产生 M 个不同的聚类结果;其次解决 M 个聚类结果的标签不一致性问题;最后通过 Co-association 矩阵对样本进行归类,得到最终的结果。

3.1　产生聚类成员

文中采用模糊 C 均值聚类(FCM)算法[11]。

但是该算法必须预定分类的个数 C 和预定的聚类中心,故本节提出一种改进的模糊聚类算法,通过取合理的聚类个数和初始聚类中心[12](初始化聚类算法),具体描述如下:

(1) 初始化空聚类集合。

(2) 对数据集中的数据进行标准化处理。

(3) 计算每两个数据的距离,得出其平均距离 $\bar{d}$。

(4) 读入第一个数据对象,并把其存入第一个聚类集合。

(5) 读入一个新的数据对象 x_j,并求其与各个现存聚类中心的距离,取其中最小的距离 $d_{\min}=\min(x_j,c_i)$,其中 $i=1,2,\cdots,k$ 为当前的聚类集合个数。

(6) 若 $d_{\min} < d$,d 为事先定义的距离,可在区间 $[0.5\bar{d},3\bar{d}]$ 之间取值(可以适当扩大其范围)。则将该数据对象加入到类 c_i 并更新该类的聚类中心。若 $d_{\min} > d$,则以该数据对象为一个新的聚类中心建立一个

新的聚类,并将 k 的值加 1。

(7) 若是最一个数据对象则算法结束,否则转(4)。

通过上面的算法,得到了一个较为合适的聚类数目和初始的聚类中心。

改进的模糊聚类算法如下所述:

1. 使用上述算法,得出初始聚类数目 c 和初始聚类中心。并给出迭代阈值 ε。

2. 在(0,1)之间随机初始化隶属矩阵 $\boldsymbol{U}$,并使 $\sum_{j=1}^{n} u_{ij} = 1$。其中 $j=1,2,3\cdots n$。

3. 用公式 $c_i = \sum_{j=1}^{n} u_{ij}^m x_j / \sum_{j=1}^{n} u_{ij}^m$ 计算聚类的 K 个聚类中心 c_i,其中 $i=1,2\cdots k$。

4. 用公式 $u_{ij} = 1/\sum_{k=1}^{c} (\frac{d_{ik}}{d_{kj}})^{2/(m-1)}$ 更新隶属矩阵 $\boldsymbol{U}$。

5. 计算公式 $J_m(u,c) = \sum_{i=1}^{N}\sum_{j=1}^{k} u_{ij}^m d_{ij}^2(X_i, C_j)$ 的函数值。如果它小于一个特定的阈值 ε 或迭代次数大于 M 则算法终止,输出聚类结果;否则,返回 2。

3.2 通过 Co-association 实现模糊聚类融合算法

假如运行 n 次 FCM 算法,每次随机选取 K 个中心点,得到 N 次不同的结果 $C_1, C_2, \cdots C_N$, 其中 $C_1 = \{P_{11}, P_{12}, \cdots P_{1K}\}$, $C_2 = \{P_{21}, P_{22}, \cdots P_{2K}\}$, $\cdots$, $C_n = \{P_{n1}, P_{n2}, \cdots P_{nK}\}$。设置一个 Co-Association 矩阵表示,任何两个数据在 N 次不同聚类过程中在同一个类中的概率,其中 $C(i,j)$ 表示第 i 个数据和第 j 个数据 在同一个类的概率。当 $C(i,j)$ 大于某一个阈值(δ) 时,将第 i 个数据和第 j 个数据并为一个聚类,剩下的数据并为一个聚类。具体算法如下:

1. 运行 n 次。

1.1 在 $[0.5\bar{d}, 3\bar{d}]$ 随机选择一个数,运行初始化聚类算法得到初始化的聚类数目和聚类中心

1.2 按上步得到的聚类数目和聚类中心运行 FCM 算法,并得到聚类结果 C_i.

1.3 更新 Co-Association 矩阵,如果第 i 个数据和第 j 个数据在同一个聚类中,那么 $C(i,j) = C(i,j) + 1/n$。

2. 使用单链接技术扫描 Co-Association 矩阵;如果 $C(i,j) > \delta$ 则数据 i 和数据 j 合并到一个聚类,直到不再存在 $C(i,j) > \delta$。将剩下的数据并为一类。

3. 输出结果。

4 实验结果与分析

文中算法中 n 和 δ 的取值对算法的性能影响很大,根据文献[13],取 $\delta = 0.6$, k 在[2,200] 取随机数,$n = 20$,效果良好。为了验证文中的算法,选用 KDD Cup 1999。该数据集由麻省理工学院 Lincloln 实验室仿真美国空军局域网环境而建立的网络流量测试数据集。数据集包含了 7 个星期网络流量,大约 500 万条连接记录。本次实验从 KDD CUP1999 中随机生成四组数据集,每个数据集包括 2100 个记录其中各种攻击记录总共 51 条,占全部记录的 2.429%,正常记录 2049 条。其中第一组数据,只包含 DoS 攻击;第二组只包括 Probing 攻击,第三组只包括 R2l 攻击,第四组包括(DOS,Probing,R2L,U2R)四种攻击。具体数据如表 1 所示。

表 1 KDD CUP1999 数据集

组数	攻击数量	攻击概率
1(DoS)	51	2.429%
2(Probing)	51	2.429%
3(R2L)	51	2.429%
4(四种攻击)	51	2.429%

用第一组(DoS)数据集,运行文中算法(C-EFC)以及 K-Means[11] 算法和 FCM 算法得出其检测率和误检率如表 2 所示。

表 2 三种算法在第一组数据上的检测结果

算法	检测率(DR)%	误报率(FPR)%
K-Means	74.9	0.67
FCM	81.5	0.78
C-EFC	91.4	0.65

用第二组(Probing)数据集,运行文中算法(C-EFC)以及 K-Means 算法和 FCM 算法得出其检测率和误检率如表 3 所示。

表 3 三种算法在第二组数据上的检测结果

算法	检测率%	误报率%
K-Means	74.7	0.72
FCM	82.7	0.74
C-EFC	89.3	0.69

用第三组(R2L)数据集,运行文中算法(C-EFC)以及 K-Means 算法和 FCM 算法得出其检测率和误检率如表 4 所示。

表 4 三种算法在第三组数据上的检测结果

算法	检测率%	误报率%
K-Means	53.8	0.65
FCM	69.8	0.69
C-EFC	88.7	0.60

用第四组(四种攻击)数据集,运行文中算法(C-EFC)以及 K-Means 算法和 FCM 算法得出其检测率和误检率如表 5 所示。

表 5 三种算法在第四组数据上的检测结果

算法	检测率%	误报率%
K-Means	68.7	0.68
FCM	77.6	0.71
C-EFC	89.4	0.67

文中算法(C-EFC)中，K_{max} 和 δ 对算法的影响很大，经过实验作者得出如下结论，当 K_{max} 大于 140 时，文中算法，检测率(DR)和误报率(FPR)的变化相对平缓，且此时有较高的检测率和较低的误报率。

5 结束语

文中采用模糊聚类融合算法，比单一的聚类算法有更好的分类效果。

通过实验可知，本文算法可以显著地提高入侵检测系统的检测率，并降低其误检率。但是文中算法的时间复杂度较大，首先要运行 n 次 *FCM* 算法其时间复杂度为 $O(n.N^2)$，另外扫描 Co-Association 矩阵的时间复杂度为 $O(N^2)$。故今后的主要研究方向为减少其时间复杂度。

参考文献：

[1] Denning D E. An Intrusion-Detection Model[J]. IEEE TRANSACTIONS ON SOFTWARE ENGINEERING,1987,13(2):222-232.

[2] Han Jiawei, Kamber M. 数据挖掘概念与技术[M]. 范明,孟小峰,等译. 北京:机械工业出版社, 2005.

[3] 薛静锋,宁宇鹏,阎 慧. 入侵检测技术[M]. 北京:机械工业出版社,2004.

[4] Popescu L. Supporting Multimedia session Mobility using SIP [C]//Communication Networks and Services Research Conference 2003. Moncton, New Brunswick, Canada: [s. n], 2003.

[5] 蒋盛益. 基于投票机制的融合聚类算法[J]. 小型微型计算机系统,2007(2):306~309.

[6] 秦 锋,陈奇明,程泽凯. 聚类融合算法研究[J]. 计算机技术与发展,2010,20(7):2-3.

[7] Strehl A, Ghosh J. Cluster Ensembles A Knowledge Reuse Frame-work for Combining Multiple Partitions[J]. Joumal of Machine Learning Research, 2003,3(3): 583-617.

[8] Topchy A, Jain A K, Punch W. A Mixture Model for Clustering Ensembles[C]// Proceedings of the 4th SIAM International Conference on Data Mining. [s. l.]:[s. n.],2004:379-390.

[9] Minaei-Bidgoli B, Topchy A, Punch W F. *A Comparison of* Resampling Methods for Clustering Ensembles[C]// Intl Conf on Machine Learning Models, Technologies and Applications (MLMTA 2004). [s. l.]:[s. n.],2004:939-945.

[10] 崔 阳,杨炳儒. 超图在数据挖掘领域中的几个应用[J]. 计算机科学,2010(6):2-3.

[11] Cannon R L, Dave J, Bezdek J C. Efficient Implementation of the Fuzzy C-Means Clustering Algorithms[J]. Pattern Ayalysis and Machine Intelligence,1986(10):213-215.

[12] 蒋盛益. 基于投票机制的融合聚类算法[J]. 小型微型计算机系统,2007(2):34-36.

[13] Fred A, Jain A K. Data Clustering Using Evidence Accumulation[C]//Proceedings of the 16th International Conference on Pattern Recognition(ICPR 2002). [s. l.]:[s. n.],2002:276-280.

(上接第 249 页)

其是基于样式表合并的信息隐藏方案，将不同的信息包含在不同的样式表当中，只有将所有样式表关联才能够得到完整的隐藏信息，充分体现了密码学中密钥分存的思想。由于 XML 文档存储与显示分离的特性，以上方法实现简单但是具有很好的安全性、隐蔽性和应用价值。

参考文献：

[1] 桂 浩,陈 刚,范 昊. XML 开发技术教程[M]. 武汉:武汉大学出版社,2008.

[2] 秦振海,谭守标,徐 超. 基于 Web 的表格信息抽取研究[J]. 计算机技术与发展,2010,20(2):217-220.

[3] Voyatizis G, Pitas I. The use of watermarks in the protection of digital multi media products[J]. Proceedings of IEEE, 2004,87(7):1197-1207.

[4] Inous S, Makino K, Murase I, et al. A proposal on information hiding methods using Xml[EB/OL]. [2008-10-25]. http://takizawa. ne. jp/nlp_xml. pdf.

[5] 吴 晶,王书文. 基于 XML 语言的信息隐藏方法[J]. 中国安全科学学报, 2005,15(12):78-82.

[6] 周 莉,王炼红,李丽娟. 一种基于 XML 文档的数字水印方案[J]. 湖南大学学报(自然科学版),2007,34(5):83-86.

[7] 蒋 斌,平西建,张 涛. 基于 XML 伪编译的信息隐藏算法[J]. 东南大学学报(自然科学版),2007,37(9):173-176.

[8] Sion R, Atallah M, Prabhakar S. Resilient information hiding for abstract semi-structures[C]//Proceedings of the Second Workshop on Digital Watermarking. Seoul: Springer-Verlag, 2004:141-153.

[9] Amblard D G. Query-preserving watermarking of relational databases and xml documents[C]//Proceeding of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on principles of database. systems. California, USA: ACM Press, 2003:191-201.

[10] Sperberg-McQueen C M, Burnard L. A Gentle Introduction to SGML[EB/OL]. [2011-03-02]. http://www. isgmlug. org/sgmlhelp/g-index. htm.

[11] 孙更新,肖 冰,彭玉忠. XML 编程与应用教程[M]. 北京:清华大学出版社,2010.

[12] Naor M, Shamir A. Visual Cryptography[C]//Advances in Cryptography- *EURO-CRYPT'94*. Perugia, Italy: [s. n.], 1994:1-12.