改进 K-means 算法在网络入侵检测中的应用

张宏博，江 弋
（厦门大学信息科学与技术学院计算机系 福建 厦门 361005）

【摘 要】：为了增强检测非法入侵的能力，把数据挖掘中的相关算法作用于入侵检测。介绍了入侵检测的概念和 k-means 算法。针对 k-means 算法的缺点，提出应用于入侵检测的改进 k-means 算法。通过实验，改进的 k-means 算法能得到比较好的检测率。

【关键词】：网络安全；入侵检测；k-means

1、引言

随着网络技术的蓬勃发展，人们在受益于网络的同时，网络也成为许多不法分子攻击的目标。网络入侵检测应用而生。它是一种主动保护网络安全的技术，能够及时发现系统中的异常行为，提醒用户注意和防范。文中将数据挖掘应用到网络入侵检测当中，能够检测出非法行为，从而保护了网络的安全性。

2、入侵检测

2.1 入侵检测的定义

入侵检测是把网络中的非法入侵检测出来。它能从计算机网络或系统中的关键点收集信息并分析，发现非法的攻击行为，并在系统中做出响应。

2.2 入侵检测的类型

入侵检测系统根据不同的数据来源可以分为：基于主机的入侵检测系统和基于网络的入侵检测系统。前者主要是检测电脑用户在主机上的行为。基于网络的入侵检测系统主要检测网络上的攻击行为。

入侵检测系统根据不同的检测角度可分为：异常检测（Anomaly detection）和误用检测（Misuse detection）。异常检测是将检测数据和正常行为做比较，判断是否是非法入侵。而误用检测是将检测数据和一个事先准备好的具有攻击特征的特征库做比较，判断是否是入侵。后者是很多入侵检测系统采用的方法。

随着科技的迅速发展，网络中的信息量增长的越来越快，存储的信息量也越来越庞大，很多入侵检测系统不能对大量的数据进行有效的分析，从而造成了检测效率的低下，而数据挖掘工具能够克服以上的缺点，因此，在网络入侵中运用数据挖掘方法，可以得到较好的检测结果。

3、k-means 聚类

数据挖掘是从海量的数据中抽取出潜在的、有价值的知识(模型或规则)的过程。它是一个利用各种分析工具在海量数据中发现模型和数据间关系的过程，这些模型和关系可以用来做出预测。数据挖掘任务一般可以分 2 类：描述和预测。描述性挖掘任务刻画数据库中的一般特性；预测性挖掘任务从当前数据上进行推断，以进行预测。

数据挖掘可以分为三个主要的阶段：数据准备、数据挖掘、结果的评价和表达。数据库中的知识发现是一个多步骤的处理过程，也就是这三个阶段的反复过程。目前所用到的数据挖掘方法主要有统计方法，关联规则，聚类分析，决策树方法，神经网络，粗糙集，支持向量机等。本文主要用到了 k-means 聚类算法和它的改进算法。

3.1 k-means 聚类算法

k-means 算法，是一种使用广泛的聚类算法。该算法将要生成的簇的数目 k 作为输入的参数，将所有的数据样本分为 k 组，同一组的数据的相似度比较高，而不同组之间的相似度比较低。这样，就将数据样本做了分类。算法首先随机的选择 k 个对象作为初始聚类中心，对于其它的数据对象，计算它和每个中心的距离，一般是计算欧式距离，然后把该对象分到距离它最近的那个组当中。再重新获得每个组的平均值，作为该组的新的聚类中心，一直重复以上的步骤，当准则函数收敛后停止。

传统的 k-means 算法存在以下缺点：

（1）k-means 算法中聚类个数 k 需要预先给出来。对于一组未知的数据，要事先确定分为几组，的确是一件比较困难的事，往往是根据经验来给定 k 的值，而这样会导致检测结果不准确。

（2）k-means 算法里，初始聚类中心的选择是很重要的，不同的初始聚类中心，可能会使聚类结果有很大得不同，该算法中，初始聚类中心是随机选择的，可能会导致检测结果不尽人意。

（3）该算法只有在簇的平均值被定义的情况下才能使用。

3.2 k-means 算法的改进

因为 k-means 算法存在不足，所以对初始聚类中心的选择，簇中心平均值的计算等方面做出改进，从而在一定程度上改善了聚类的结果。

3.2.1 改进初始聚类中心

在数据空间里，低密度的空间对象区域分割高密度的空间对象区域，把处于低密度区域的数据点叫做噪声。我们取相互之间的距离最远的 k 个点作为初始的聚类中心，这 k 个点都要位于高密度，这样可以防止将噪声点取到。

这里定义密度参数的概念，是用来求得对象 Xi 所在的密度，判断它是在高密度区域还是在低密度区域。将数据对象 Xi 作为中心，密度参数定义为包含常数 M 个对象的半径。用 t 表示，计算出来的 t 值越大，说明该数据点位于低密度区域中，计算出来的 t 值越小，说明该数据点位于高密度区域中。通过上述的方法，把每个数据对象的密度参数计算出来，得到所有的高密度区域的点，把这些点组成一个集合 D.样本点与一个样本集中所有样本点当中最近的距离被称作该样本集和该样本点的距离。

集合 D 里都是位于高密度区域的点，在 D 中，将具有最高密度的对象最为第一个初始中心点，记作 W1，然后取距离 W1 最远的集合 D 中的点作为第二个初始中心点 W2，求集合 D 中各个对象 Xi 和 W1，W2 的距离 dist(Xi,W1),dist(Xi,W2),然后取距离最远的对象作为第三个初始中心点 W3，也就是满足 MAX(MIN(dist(Xi, W1), dist(Xi, W2)))(i=1,2,……,n)，根据这个方法，知道找到最初的 K 个初始聚类中心，这样方法找到的初始聚类中心，都是位于高密度区域中得点，并且相互距离都是最远的。

具体的流程如下：

①求每个数据对象和其它数据对象之间的距离，这里采用欧式距离。

②通过上述密度的定义，求得每个数据对象的密度参数，把位于的高密度区域中的数据放在集合 D 中，低密度区域中的数据是无效的，这里删掉不使用。

③在集合 D 中，找到位于最高密度区域的数据，将他作为第一个初始中心 W1，并将 W1 放到集合 W 中，在集合 D 中删掉这个数据。

④在 D 中，寻找离集合 W 最远的数据，作为第二个初始中心 W2，并将 W2 放到集合 W 中，在集合 D 中删掉这个数据。

⑤继续以上的步骤，知道 K 个初始中心全部找到，也就是集合 W 中已经有了 K 个数据。

3.2.2 通过对特征值加权改进算法

在含有 n 个数据对象的数据集合中，每个数据对象对于知识发现的作用是不同的，为了区分这些相异之处，给每个数据对象赋予一个权值。在这里采用 Domeniconi 等人提出的权重设置方法。该方法的基本原理是对类内分布一致性好的特征赋予较大的权重，在不同的类内相同的特征赋予不同的权重。类内分布的一致性主要通过类内该特征的方差大小度量。

假设X代表整个数据集，X_i代表第i类数据集，x代表数据对象，E_{ir}代表特征r在第i类的分布方差，w_{ir}代表特征r在第i类的权重，c_k代表第k类的类中心向量。

$$X_i = \{x | i = \arg\min_k dist_w(c_k, x)\} \quad (1)$$

$$dist_w(c_k, x) = [\sum_{j=1}^{d} w_{kj}(c_{kj} - x_j)^2]^{1/2} \quad (2)$$

$$E_{ir} = [\sum_{x \in X_i} (c_{ir} - x_r)^2]^{1/2} / |X_i| \quad (3)$$

其中c_{ir}表示第i类中心的第r特征值，x_j表示数据对象x的第j特征，$|X_i|$表示第i个数据集X_i的数据对象个数，第i类的第r特征的权重w_{ir}定义如下：

$$w_{ir} = \exp(-h * E_{ir}) / (\sum_{k=1}^{d} \exp(-2h * E_{ik}))^{1/2} \quad (4)$$

h是正的常数，在这里，h定义为12。

为了避免由于E_{ir}过大而导致$\exp(-h*E_{ir})$由于计算精度问题而趋于零值，需要预先对数据对象进行标准化处理，经过实验，发现令$x_j = x_j / \mu_{x_j}$可以取得较好的结果，其中，μ_{x_j}为x_j的均值。

基于上述方法对特征权重重新计算，基于调整后的特征权重重新计算各样本点与各聚类中心的距离，并在此基础上重新分配给样本点，计算新的聚类中心。

综上所述，改进后的算法流程如下：

①用3.2.1的方法的到K-means算法的K个初始中心点。

②最初的权值是 w_{ir}=1/d ，d是代表数据的维数。

③按照公式(1)、(2)把各个数据对象划分到相应的数据对象集合X_i中。根据公式(3)、(4)计算新的权重系数 w_{ir}。

④根据公式(1)、(2)重新计算各个数据对象与当前各聚类中心的距离，将数据对象划分到相应的数据对象集合X_i中。

⑤重新计算各聚类中心。

⑥重复步骤33、44、55直到算法收敛或达到指定的迭代次数。

4、实验结果与分析

文中采用KDD Cup 99入侵数据包进行实验。KDD Cup 99是美国麻省理工学院仿真美国空军局域网环境而建立的测试数据集。数据集中得记录大约有500万条。这500万条记录模拟了多种网络环境下的入侵，每条记录有42属性，并且已经给出是正常数据还是攻击数据。

4.1测试数据的归一化处理

入侵数据包中的每条记录如下所示：

1,tcp,http,SF,159,459,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,32,33,0,0,0,0,1,0,0.06,32,255,1,0,0.03,0.02,0,0,0,0,nor-

mal

在每条记录中，包括了协议，标记，持续时间，接收字节数，发送字节数等属性。在实验中我们选择其中15个关键的数值属性进行聚类。对这些数值型数据首先要归一化。

4.2实验结果

本算法是在windows xp下的Visual C++6.0环境下实现的。从入侵数据包中选取1000条正常数据和100条攻击数据作为测试数据集M1。选取10000条正常数据和1000条攻击数据作为测试数据集M2。聚类个数为5。结果分为5类，一类是正常数据类，其它四类是攻击数据类，统计结果如下表。

	改进后的算法		k-means 算法	
	M1	M2	M1	M2
AA	93	972	93	966
AN	7	28	7	34
攻击数据检测率	93%	97.2%	93%	96.6%
NA	8	178	19	520
NN	992	9822	981	9480
正常数据检测率	99.2%	98.22%	98.1%	94.8%
总检测率	96.1%	97.71%	95.55%	95.7%

AA：攻击数据被检测成攻击数据的数量

AN：攻击数据被检测成正常数据的数量

NA：正常数据被检测成攻击数据的数量

NN：正常数据被检测成正常数据的数量

5、结论

随着应用软件和操作系统的复杂化，网络安全受到越来越多的威胁。将数据挖掘方法引入到网络入侵检测，有利于从大量数据中发现攻击行为，保护网络的安全。在传统的k-means基础上，本文应用改进的k-means算法测试网络攻击数据，对于检测率有了一定的提高。

参考文献：

[1] 张杰，戴英侠，入侵检测系统技术现状及其发展趋势[J].计算机与通信，2002(6):28-32.

[2] 唐正军，李建华. 入侵检测技术 [M]. 北京：清华大学出版社，2004.

[3] 汉德.数据挖掘原理[M].北京:机械工业出版社，2003.1-2

[4] Domeniconi C , Papadopoulos D , Gunopulos D , Ma S1 Subspace Clustering of High Dimensional Data In : Proc. of the Fourth SI- AM Intl. Conf. on Data Mining ,2004. 517~521.

[5] Ren Jiangtao, Shi Xiaoxiao, Sun Jinhao. An Improved K-Means Clustering Algorithm Based on Feature Weighting. Computer Science. 2006Vol133, No 17. 186~187.

[6] Huang Zhexue. Extensions to t he K-Means Algorithm for Clustering LargeData Sets with Categorical Values. Data Mining and Knowledge Discovery ,1998. 283~304.

(上接第 98 页)

捷，基于集成技术建设校园信息系统，对于整合优化教学资源、提升辅助技能教学能力、拓宽技能教育培训层次和范围、建立市场导向型技能人才队伍培训基地，对于我省技工院校发挥全国技工教育排头兵示范效应、实现技工教育改革创新发展、实施人才强国战略具有重要意义。

参考文献：

[1]李志华，陈晓宁，郭玉娇(2009)。〈应用系统集成的研究和应用〉，3 期，97-100。化工高等教育。

[2]杨延娇(2006)。〈浅谈企业应用集成技术〉，《管理科学》35 期，99。

[3]吴文哲(2010)。〈基于 SOA 架构的高校应用系统集成方案设计〉，《信息技术》9 期，13-14。

[4]朱迅，丁勤，杨丽波(2010)。〈基于 SOA 的数字化校园信息系统集成研究〉，《信息化研究》，36 期，59-61。

[5]王全旺，赵兵川(2008)。〈我国高级技工学校教育信息化现状及提升途径〉，2 期，12-13。中国职业技术教育。

[6]张艳琼，蔡瑞英(2009)。〈基于 Web 服务的数字校园应用集成研究〉，《微处理机》，3 期，67-69。

[7]孙昌爱，金茂忠，刘超(2002)。〈软件体系结构研究综述〉，《软件学报》，13 期，1228- 1237 。

[8]李润洲，宋彩利(2006)。〈校园网格数据集成中间件体系结构研究〉，《西安科技大学学报》，4 期，532-535。

[9]EMMERICH W，ELLMER E，FIEGLEIN H. TIGRA(2001)。〈an architectural style for enterprise application integration〉，《Proc of the 23rd International Conference on Software Engineering》，567- 576。

[10]任午令，唐任仲，郭尚鸿，刘永清(2007)。〈基于构件的企业应用集成技术〉，《浙江大学学报》，41 期，1283-1287。

[11]李强，南理勇(2009)。〈基于 Web Services 的企业应用集成方案〉，《软件导刊》，8 期，124-125。

[12]宋庭新，黄必清，魏春梅(2008)。〈基于语义 Web 服务的协同物流与集成技术研究〉，《计算机集成制造系统》，14 期，588-594。

[13]麻丽娜，苑津莎，李新叶(2005)。〈基于 Web Service 的电力企业应用集成技术研究与实现〉，《电力系统通信》，26 期，40-42。

[14]胡建理，王嘉祯(2005)。〈基于 Web Service 的电力企业应用集成技术〉，《计算机工程与设计》，26 期，2634-2637。

[15]杨林(2007)。〈基于企业服务总线的企业应用集成技术研究〉5 期。华中科技大学。

[16]蒋丽丽(2009)。〈企业应用集成技术研究进展〉，《福建电脑》，11 期，42-43。