10.3969/j.issn.1000-0755.2016.11.010

基于数据挖掘的网络异常检测方法的研究

宋先强　高仲合　刘 泷　国凯平
（曲阜师范大学信息科学与工程学院，山东日照，276800）

摘 要：为提高 K-means 聚类算法在异常检测中的效果，以聚类分析为主线，针对传统的K_means聚类算法在初始聚类中心点选择随机性和K值预先设定的问题，提出了一种改进的K_means聚类分析算法，算法引入密度参数和距离理论。依据密度理论和最大距离找出k个初始化中心点。并对算法进行仿真实验，实验证明，新的算法具有良好的效果。

关键词：数据挖掘；聚类；k-means聚类算法；异常检测

Research on Network Anomaly Detection Method Based on Data Mining

Song Xianqiang　Gao Zhonghe　Liu Shuang　Guo Kaiping
（School of Information Science and Engineering，Qufu Normal University, Rizhao,276800,China)

Abstract：In order to improve K-means clustering algorithm effect in Anomaly Detection,use Cluster analysis to the main line,for K_means clustering algorithms the problems in the initial cluster centers select random and the preset value K. an algorithm to calculate the number of the Cluster Center was given.Proposed an improved K-means Clustering Algorithm,we used Density parameter and the theoretical of distance in this algorithm. according to the theoretical density and maximum distances to find k initialization center. For This Algorithm ,we set to test, The results show the algorithm have a higher detection rate and a lower false alarm rate.

0 引言

数据挖掘[1]是从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中，发现并提取隐含在其中未知的、可信的、有用的模式的过程。聚类是数据挖掘中的一类重要技术，是分析数据并从中发现有用信息的一种有效手段。它将数据对象分组成为多个类或簇，使得在同一个簇中的对象之间具有较高的相似度，而不同簇中的对象差别很大。聚类从数学分析的角度提供了一种准确、细致的分析工具。异常检测过程即是数据分析的过程。因此，在入侵检测时，可利用数据挖掘技术，对海量网络数据分析处理，从中提取尽可能多的隐藏的安全信息。本文在数据挖掘和异常检测理论的基础上，引入密度参数和距离理论，提出了新的确定k个初始聚类中心点的方法和去除孤立点的方法。最后本文使用KDD CUP99数据集来测试算法的性能。结果表明，此方法具有较高的检测率和较低的误报率，达到了预期的目标。

1 聚类算法应用于异常检测相关研究

利用数据挖掘中的聚类分析方法进行异常的入侵检测时，不需要对数据对象作正常或异常标记，而是基于以下两点假设[2]：①在网络监测过程中，正常流量的数目要远远多于异常流量的数目；②异常流量与正常流量之间存在着实质的差异。聚类分析方法的基本思想就是由于异常流量和正常流量不同并且数目相对很少，因此对网络流量特征进行聚类分析，挖掘出其中表现异常的数据流，结合入侵检测模型判别其为何种入侵。

在聚类算法中，对 K-means 算法的改进研究最为广泛。K-means是一种常用在异常检测中的聚类分析方法，它是J.B.MacQueen在1967年提出的一种基于划分的动态聚类算法。但是，传统的K-means算法应用在异常检测中时却存在如下问题[3]：

（1）不能聚类前确定聚类个数k值
（2）初始聚类中心的选取影响聚类的结果
（3）孤立点对聚类结果的影响

2 优化初始中心的K-means算法

2.1 对孤立点的处理

以往研究者通常是用距离大小判定是否为孤立点，定义如下[4]：

定义1 如果对于每个数据点i，其与其他点的距离和为D_i，各个点的距离均和为H，如果D_i>H，则认为该点是孤立点。

其中$D_i=\sum_{j=1}^{n}\sqrt{\sum_{h=1}^{d}(x_{ih}-x_{jh})^2}$，$H=\sum_{i=1}^{n}\frac{D_i}{n}$，n为样本数据，d为数据的维数。

这种处理方式只注重了孤立点与其它点的位置，却忽略了孤立点分布上的稀疏性。可以把所有孤立点都可以移除，但这样容易把密度边界点当成孤立点，从而影响聚类效果。

因此，有的研究者提出了聚类的密度理论，用密度来判断孤立点。点的密度定义如下[5]：

定义2 对空间中任意一点p和距离r，以p为中心，以r为半径作超维球体，落在该球体内的数据点的个数称为点p关于距离r的密度，记作D（p，r）。

这种方法也是比较片面的，这样容易把一些密度较小的点当成孤立点进行处理，从而影响聚类效果。我们应该把孤立点与密度边界点区别开。为此我们把距离和密度两者结合起来，当对象的密度参数足够小时，且其与类其它点距离大于均值H既可认为该对象为孤立点。

密度边界对象点是指其密度参数小于制定密度指数Q，并且包含于聚类中心p的r半径超维球状范围里，既其与中心对象距离小于类中所有对象与中心点的距离的均值。密度边界对象点的特点是距离类的中心近，但是密度值小。依据改进后的方法可以将密度边界点与孤立点区别开。

密度半径r密度指数Q的取值方法如下：

在混合类型数据的样本集中(包含正常数据和各种攻击数据的混合样本集)，把平均欧式距离较小的类型数据的平均欧式距离作为r的初始取值，在此基础上实验得到最佳取值。

3.2 k值的确定

K-means 算法在进行聚类之前要求知道 K 值，这对于没有经验的用户来说是比较困难的。K-means 算法通过欧氏距离来划分聚类的，假设精确的最佳聚类个数为k，若初始值选择k_1，并且k1< k，则说明至少有两个合理划分的类被归结成了一类。若初始值选择k_2>k，则说明至少有一个合理划分的类被再次划分成了若干个类[6]。虽然精确的最佳聚类个数值（k）难以确定 ，但却可以通过得到聚类个数的上限(K)来缩小聚类个数 K 的设置范围。很多研究者使用的经验规则[7]为：

$K = \lfloor \sqrt{N} \rfloor$或$\lfloor lnN \rfloor$其中N为样本总数

3.3 选取适当的初始聚类中心

本文提出了一种新的寻找k个聚类中心的方法，从而可以得到更好的划分效果。

算法的基本思想是：给定一个密度半径r，按照点的密度的定义，计算每一个数据对象关于距离阈值的密度，把对象按照密度排序，尽量选中密度较大的且相对距离较远的k个对象作为初始聚类中心。

初始聚类中心选择算法：

输入：有N条记录的数据集A′，密度半径r；

输出：初始聚类中心集合S

（1）初始化初始聚类中心集合S为空

（2）for（读取数据集A′中的每一个记录P）

计算该记录的点密度；

(3）按照点密度大小对集合D进行降序排序；

（4）Set（将对象集中密度参数最大的点作为第1个初始聚类中心C1，将与C1距离最大的一个高密度参数点作为第2个初始聚类中心C2）

（5） Delete（C1，C2）

（6）While（A′中仍有未读取记录）{

读取当前记录；

计算当前记录与S中每个初始聚类中心的距离；

把其中的最小距离保存在 中；

if（d_{min}>S中记录间的最小距离）

if（S中记录个数<k）

把当前记录加入到S中；

else if (S中记录个数= k)

{ 计算s中距离最小值点对象中心点为新的初始中心点，delete S 中距离最小值点对记录，把当前记录加入到S中（对象C1和C2不进行计算）；

}

}

3.4 基于改进的k-means的异常检测算法

结合以上研究，给出改进的k-means算法流程如下：

输入：密度半径r，n条记录的数据集A。

输出：k个聚类

（1）扫描一次数据集A，For（读取样本集合A中的每一个数据i）。

（2）计算每个数据的点密度，并计算出每个数据点的距离和D_i与距离均和H。

（3）如果数据对象密度值小于密度值Q，并且D> H则视为该对象为孤立点t。

（4）去除A中的孤立点数据，得到新的数据集A'，并记录A'中的样本个数m=n-t，输出孤立点。

（5）运行改进的获取初始聚类中心的算法，获得k个初始聚类中心。

（6）运行传统k-means算法进行聚类分析。

4 实验结果与分析

建立入侵检测模型，选用KDD Cup99实验数据[8]。对原始的KDD Cup99数据集中很多异常数据过滤掉，

使得入侵数据占总数据量的1%~2%。

本文中把数据挖掘算法中检测率和误检率两个值作为改进的K-means算法性能的度量标准。

$$检测率 = \frac{检测到的异常数据数目总和}{数据集中异常数据总数} X\ 100\%$$

$$误检率 = \frac{被误认为是异常的正常数据数}{数据集中正常数据总数} X\ 100\%$$

本次仿真实验，我们对KDD Cup99数据集进行了提取，随机选取10000条已知正常数据和异常数据比例的数据，用来确定密度半径r和Q的值。然后使用5组5050个样本数据集，每个数据子集中正常数据个数约占总数的99%，同时过滤选取四种入侵活动数据均匀分配到这5个数据子集中。$K=\lfloor\sqrt{N}\rfloor$ 确定k的大小，实验结果如表1所示。

表1 引入密度和距离中心理论改进的K-means算法的检测结果

样本集	正常数据量	K值大小	异常数据量	检测率（%）	误检率（%）
数据集1	5000	71	50	88.13	8.41
数据集2	5000	71	50	87.32	9.64
数据集3	5000	71	50	92.41	5.36
数据集4	5000	71	50	85.41	6.15
数据集5	5000	71	50	91.62	3.42

表2 传统k-means算法的检测结果

样本集	正常数据	K值大小	异常数据	检测率	误检率
数据集1	5000	30	50	78.42	15.51
数据集2	5000	40	50	79.63	11.32
数据集3	5000	50	50	82.51	10.58
数据集4	5000	60	50	84.65	9.14
数据集5	5000	70	50	88.69	8.62

通过上面的数据分析可以看到，利用该方法进行异常入侵检测可以有效地将正常数据和攻击数据区分开来，通过表1和表2的结果比较可知，传统的k-means算法当k值接近$K=\lfloor\sqrt{N}\rfloor$时，其检测率也在慢慢提升。改进的k-means算法比传统k-means算法在检测率上提升了不少，误检率也比k-means算法降低不少，达到了预期的目标。

5 结束语

文中结合了数据挖据中聚类算法在异常入侵检测中的特点和优势，引入密度参数和距离概念对现有的K-Means算法进行了改进，并利用KDD Cup99实验数据进行仿真实验，其结果表明该算法能有效提高检测的效率，较之传统k-means算法在误检率上有很大提高。

本方法在实践应用中还需根据实际情况做进一步改进，降低算法的时间复杂度比较高，以提高性能，使之处理大规模数据时提高效率。

参考文献：

[1] 郭春. 基于数据挖掘的网络入侵检测关键技术研究[D].北京邮电大学,2014.
[2] 李涵.一种改进的聚类方法在异常检测中的应用[J].微电子学与计算机,2010,08:66-69.
[3] Hansen P, Mladenovic N. J- means: a new local search heuristic for minimum sum- of- squares clustering[J]. Pattern Recognition,2002,34(2): 405- 413.
[4] 熊忠阳,陈若田,张玉芳. 一种有效的K-means聚类中心初始化方法[J].计算机应用研究,2011,11: 4188-4190.
[5] 贾永娟. 基于密度的改进K-Means文本聚类算法研究[D].山西师范大学,2014.
[6] Using an Improved Clustering Method to Detect Anomaly Activities[J]. Wuhan University Journal of Natural Sciences,2006,06:1814-1818.
[7] 于剑，程乾生．模糊聚类方法中的最佳聚类数的搜索范围[J]. 中国科学, 2002, 32(2) :274-280.
[8] Li Yang, Fang Bin-xing, Guo Li,et al. A Network Anomaly Detection Method Based on Transduction Scheme [J]. Journal of sofware, 2007,18(10): 2595-2604

作者介绍

宋先强(1991-)，男，硕士研究生，主要研究方向：计算机网络
电话：13793426416
电子信箱：940847123@qq.com
联系地址：山东省日照市烟台路80号 曲阜师范大学 信息科学与工程学院
邮政编码：276800
高仲合（1961-），男，教授，硕士生导师，主要研究方向：计算机网络与通信
刘泷（1990-），男，硕士研究生，主要研究方向：网络信息安全
国凯平（1992-）男，硕士研究生，主要研究方向：网络信息安全