PCA-AKM算法及其在入侵检测中的应用

牛 雷 孙忠林

(山东科技大学计算机科学与工程学院 山东 青岛 266590)

摘 要 初始聚类中心是指在聚类的过程中首次被选为中心的点或对象。针对传统的K-means算法由于随机选择初始聚类中心而造成的聚类结果不稳定的问题,提出PCA-AKM算法。该算法利用主成分分析方法提取数据集中的主要成分,实现数据降维,使用自定义指标密权值选择初始聚类中心,避免聚类中心局部最优问题。将该算法与K-means算法在UCI数据集上进行聚类对比,其聚类稳定性高于传统K-means算法。在KDD CUP99数据集上,对所提算法进行入侵检测仿真,实验结果证明该算法检测率高,误检率低,能够有效提高入侵检测的准确率。

关键词 K均值算法,主成分分析,密权值,入侵检测

中图法分类号 TP393.08 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2018.02.039

PCA-AKM Algorithm and Its Application in Intrusion Detection System

NIU Lei SUN Zhong-lin

(College of Computer Science and Engineering, Shandong University of Science and Technology, Qingdao, Shandong 266590, China)

Abstract The initial clustering center is the point or object selected for the first time in the clustering process. Aiming at the instability of clustering results in traditional K-means algorithm caused by choosing the initial clustering centers randomly, the PCA-AKM algorithm was proposed. The algorithm uses the principal component analysis to extract the main components of the data set to achieve data dimensionality reduction, and then uses the self-defined indicators Dw to choose the initial clustering centers, avoiding the clustering center local optimum. Comparison with the K-means algorithm on the UCI data set proves that the clustering stability of the PCA-AKM algorithm is higher than that of K-means. Experiment proves that the algorithm has high detection rate and low false detection rate on KDD CUP99 data set when it is used to simulate intrusion detection, and the algorithm can improve the accuracy of intrusion detection effectively.

Keywords K-means algorithm, Principal component analysis, Dw, Intrusion detection

1 引言

聚类分析是一个把数据对象划分成子集的过程,每一个子集是一个簇,簇中对象彼此相似,簇间对象不相似。由聚类分析产生的簇的集合称为一个聚类。聚类分析已经广泛地应用于许多领域,包括商务智能、图像模式识别、Web搜索、生物学和安全等。

K-means是数据挖掘的经典聚类算法之一,但其仍存在过分依赖k值,以及随机选取初始中心会造成聚类效果不理想的问题。目前对K-means算法的改进主要从初始聚类中心的选取、k值的选取、聚类收敛条件以及处理分类属性数据4个方面进行。文献[1-7]主要从初始聚类中心的选取方面进行了研究。文献[1]提出了基于数据对象分布密度以及计算最近两点的垂直中点的方法来确定k个初始聚类中心。文献[2]提出了最大距离法来选取聚类中心并计算样本点两两之间的距离,其选取距离最远的两个样本点作为初始聚类中心。文献[3]提出了基于数据样本分布情况的动态选取初始聚类中心的算法,根据数据点的距离构造最小生成树,并对最小生成树进行剪枝,得到k个初始数据集合,进而得到初始聚类中心。文献[4]提出了基于距离阈值选取初始聚类中心的方法,它计算样本集的均值并将其作为第一个初始中心,然后根据最大最小距离算法动态确定其他初始聚类中心。文献[5]提出了粒子群算法,将每个类作为一个粒子群,计算每个粒子的适应度值,并由适应度值来确定初始聚类中心。文献[6]提出了网格法来寻找聚类中心,先进行网格均分,划分出在线网格以及离线网格,将有重叠的网格进行合并,并将网格内点的均值作为初始聚类中心。文献[7]将遗传算法和自适应权重结合后运用到K-means中,因为遗传算法具有全局搜索能力,所以利用它来寻找最优聚类中心。

本文提出基于PCA的加速K-means算法,利用自定义指标密权值选取初始聚类中心。该指标由样本密度及属性权值共同构造,使用该指标能够有效防止聚类中心局部最优问题。通过实验证明,本文算法在UCI数据集上的聚类精度高于传统的K-means算法,并且该改进算法能够提高入侵检测

到稿日期:2017-04-24 返修日期:2017-07-07 本文受"十二五"国家科技支撑计划基金项目(2014BAL04B06)资助。

牛 雷(1992-),男,硕士生,主要研究方向为数据库、数据挖掘,E-mail:514654471@qq.com;孙忠林(1962-),男,博士,教授,主要研究方向为模式识别、数据库系统、系统集成及安全工程,E-mail:zhonglinsun@163.com(通信作者)。

系统的检测率,比传统的 K-means 算法具有更好的检测效果。

2　K-means

K-means 算法将簇内点的均值作为簇的中心。首先,算法随机选择初始聚类中心,在数据集 D 中随机选择 k 个对象,将这 k 个对象作为初始聚类中心。对于剩下的对象,根据其与各个簇中心的欧氏距离[5],将其分配到最近的簇中。然后,K-means 算法开始更新簇中心,使用上一次分配到该簇的对象来计算均值,并将簇均值作为新的簇中心,之后利用新的簇中心重新分配所有的对象,不断迭代,直到分配稳定,即本轮形成的簇与前一轮形成的簇相同。

K-means 算法的具体步骤如算法 1 所示。

算法 1　K-means

输入:簇数目 k,包含 n 个对象的数据集 D

输出:k 个簇的集合

方法:

1. 从 D 中任意选择 k 个对象作为初始聚类中心;
2. Repeat
3. 　根据簇中对象的均值,将每个对象分配到最相似(即距离最近)的簇;
4. 　更新簇的均值,即重新计算每个簇中对象的均值;
5. Until 不再发生变化。

3　聚类定义

定义 1(欧几里得距离)　K-means 算法采用欧几里得距离来度量样本间的相似性[5]。

$$d(x,y)=\sqrt{(x_1-y_1)^2+\cdots+(x_n-y_n)^2} \tag{1}$$

其中,x,y 代表 n 维数据。

定义 2(聚类准则函数 E)　K-means 算法用此函数来判断是否达到最优聚类效果,E 越小则聚类效果越好[8]。

$$E=\sum_{i=1}^{k}\sum_{x\in C_i}|x-\overline{x_i}| \tag{2}$$

其中,k 为聚类个数,C_i 为 k 个聚类中的第 i 个类,$\overline{x_i}$ 为类 C_i 的聚类中心。

定义 3(最大最小距离法[9])　该方法是一种用来寻找初始聚类中心的方法,当选择第 $i(i>2)$ 个聚类中心时,先计算样本与已确定聚类中心间的距离,然后将最小的距离作为其最大最小值,选取样本中最大最小值最大的样本点作为下个聚类中心,直至找到 k 个聚类中心为止。

$$d_{\max}(x,C)=\max(\min(d(x,C))) \tag{3}$$

定义 4(数据对象分布密度[1])　数据样本总体 $D=\{x_1,x_2,\cdots,x_n\}$,数据对象 x_i 的分布密度为 $dens(x_i)$,其中 q_{ij} 表示数据对象 x_i 和 x_j 之间连接的所有路径,L 表示在路径中经过的数据点的个数,$p(p>1)$ 表示密度系数。

$$dens(x_i)=\sum_{j=1}^{n}\frac{1}{\min\limits_{q_{ij}}\sum\limits_{k=1}^{L-1}p^{d(x_k,x_{k+1})}-1} \tag{4}$$

4　基于 PCA 的加速 K-means 算法

主成分分析算法(Principal Component Analysis,PCA)的基本思想是:当涉及多维属性信息,且属性之间存在相关性时,应用主成分分析将属性重新组合成无相关性的主成分,用于表示原有信息。使用主成分分析算法可以有效地降低样本集的维数,提高运算效率。PCA-AKM 算法利用主成分分析提取多维数据中的重要属性,可以有效地提高入侵检测时算法的执行效率。它利用密权值寻找最优初始聚类中心,解决了传统K-means由于随机选取初始聚类中心而导致的聚类效果波动问题。

定义 5(聚类中心寻找指数-密权值 Dw)　Dw 由密度和权值两部分组成,构建 Dw 的目的是寻找最佳初始聚类中心,并防止聚类中心局部最优。

$$Dw=dens(x_i)\times\frac{\sum\limits_{i=1}^{n}|x_i-u_i|}{\sum\limits_{i=1,i\neq j}^{n}|u_i-u_j|} \tag{5}$$

其中,$dens(x_i)$ 表示各个样本点的样本密度,x_i 表示样本点第 i 列的值,u_i 表示样本集第 i 列的均值,$\sum\limits_{i=1}^{n}|x_i-u_i|$ 和 $\sum\limits_{i=1,i\neq j}^{n}|u_i-u_j|$ 的含义分别为属性内距离与属性间距离,二者的比值表示权重。将密权值较大的点作为初始聚类中心,能够获得较好的聚类效果。

标准化变换公式为:

$$Z_{ij}=\frac{x_{ij}-\overline{x_j}}{S_j},i=1,2,\cdots,m;j=1,2,\cdots,p \tag{6}$$

其中,$\overline{x_j}=\frac{\sum\limits_{i=1}^{m}x_{ij}}{m}$,$S_j=\frac{\sum\limits_{i=1}^{m}(x_{ij}-\overline{x_j})}{m-1}$。$x_{ij}$ 表示矩阵中的每个样本值,Z_{ij} 表示标准化矩阵,$\overline{x_j}$ 表示列均值,S_j 表示列标准差。

$$R=[r_{ij}]_{p\times p}=\frac{Z^{\mathrm{T}}Z}{m-1} \tag{7}$$

其中,$r_{ij}=\frac{\sum z_{kj}\cdot z_{kj}}{m-1},i,j=1,2,\cdots,p$;$R$ 为相关系数矩阵;x 为随机向量;p 为维数;Z^{T} 表示标准化矩阵 Z 的转置[10]。

PCA-AKM 算法的具体步骤如算法 2 所示。

算法 2　PCA-AKM

输入:数据集 $D_{m\times p}(p>n)$,簇数目 k,最大迭代次数 Max

输出:k 个簇的集合

方法:

1. $D_{m\times p}$ 按式(6)、式(7)计算相关系数矩阵 R。
2. 根据 $|R-\lambda I_p|=0$ 计算 p 个特征值 λ_i。
3. 由贡献率 $\frac{\lambda_i}{\sum\limits_{i=1}^{p}\lambda_i}$ 确定 n 个主成分,得到 $D_{m\times n}$。
4. For $D_{m\times n}$ 中的每一个对象,按照式(5)计算 Dw。
5. 将前 $\sqrt{m}$ 个密权值 Dw 对应的对象存入集合 C。
6. 将 C 中 Dw 最大的对象作为第一个初始聚类中心 C_1,将距离 C_1 最远的作为 C_2。
7. $C_{next}=d_{max}(x,C)$,共选取 k 个聚类中心。
8. Repeat
9. 　根据簇中对象的均值,将每个对象分配到最相似的簇中。
10. 　If $x\in C_i$,$d(C_i,C_j)>2d(x,C_i)$,则 x 不会归入 C_j 中。
11. 　更新簇均值,即重新计算每个簇中对象的均值。
12. Until M 收敛或达到最大迭代次数 Max。
13. 算法结束。

由密权值构造一个初始聚类中心选择集合,在该集合中

利用最大最小法 $d_{max}(x,C)$选择初始聚类中心。使用该方法选择聚类中心能够防止聚类中心局部最优问题。

当 $x\in C_i$,$d(C_i,C_j)>2d(x,C_i)$时,x 不会归入 C_j 中,此时可以加快算法的执行效率,即当前聚类中心到其他聚类中心的距离大于样本点到本类中心距离的两倍时,样本点不会归入该类。此处应用三角形定理:两边之和大于第三边,即 $d(x,C_i)+d(x,C_j)>d(C_i,C_j)$。当 $d(C_i,C_j)>2d(x,C_i)$时,$d(x,C_j)>d(x,C_i)$,故 x 不会被分到 j 类中[11]。

表 1 初始聚类中心比较

Table 1 Comparison of initial clustering center

序号	K-means 算法		PCA-AKM	
	初始中心	准确率/%	初始中心	准确率/%
1	73,88,144	56.33	57,10,128	88.33
2	50,20,119	80.25	57,10,128	88.33
3	11,20,8	63.21	57,10,128	88.33
4	99,27,64	83.66	57,10,128	88.33
5	41,8,130	82.34	57,10,128	88.33
6	67,10,121	85.77	57,10,128	88.33
7	3,88,14	79.21	57,10,128	88.33
8	54,72,68	84.58	57,10,128	88.33
9	36,80,101	82.71	57,10,128	88.33
10	124,47,138	81.75	57,10,128	88.33
平均值	—	77.98	—	88.33

对于初始聚类中心的选取,传统的 K-means 随机选取初始聚类中心,本文算法以密权值 Dw 为判定准则来选取聚类中心。为比较二者的差异,采用 UCI 数据集中的 Iris(150 个样本,4 个属性,3 个簇)进行初始聚类中心及聚类结果的比较,实验在相同的环境下进行,以避免其他因素的影响,比较结果如表 1 所列。

K-means 算法随机生成 3 个初始聚类中心,由于随机选择导致聚类结果波动较大,最大聚类准确率为 85.77%,最低准确率为 56.33%,平均准确率为 77.98%,准确率偏低。本文算法的选择聚类中心为(57,10,128),聚类结果基本保持不变,平均聚类准确率为 88.33%。与传统的 K-means 算法相比,本文算法的聚类准确率更高,聚类效果更加稳定。

在 UCI 数据集中选取 4 个数据集:yeast(1484 个样本,8 个属性,10 个簇),abalone(4177 个样本,8 个属性,29 个簇),magic(19020 个样本,11 个属性,2 个簇),skin(245057 个样本,4 个属性,2 个簇)。将本文算法分别与熊开玲等人[12]提出的基于核密度估计的 K-means 聚类优化算法(KNE-KM)、庄瑞格等人[13]提出的基于拟蒙特卡洛的 K 均值聚类算法(QMC-KM)、李敏等人[14]提出的密度峰值优化初始中心的 K-means 算法(CFSFDP-KM)进行聚类精及误差平方和的比较。这 3 种算法都对初始聚类中心的选取做了一定的改进,降低了 K-means 随机选择初始聚类中心对聚类结果的影响。实验聚类精度比较结果、误差平方和比较结果分别如表 2、表 3 所列。其中,*HIG* 表示最高聚类精度,*LOW* 表示最低聚类精度,*AVG* 表示平均聚类精度。

表 2 聚类精度比较结果/%

Table 2 Comparison results of clustering accuracy/%

数据集	聚类精度											
	PCA-AKM			KNE-KM			QMC-KM			CFSFDP-KM		
	HIG	*LOW*	*AVG*	*HIG*	*LOW*	*AVG*	*HIG*	*LOW*	*AVG*	*HIG*	*LOW*	*AVG*
yeast	79.54	79.54	79.54	78.66	71.23	74.63	71.23	70.23	70.75	69.11	61.28	67.53
abalone	82.13	82.13	82.13	80.19	73.22	76.22	82.13	78.21	81.25	68.82	65.70	67.13
magic	79.27	79.27	79.27	72.29	69.31	71.81	79.21	75.11	77.84	77.27	70.34	75.19
skin	83.61	83.61	83.61	81.71	71.19	76.29	80.91	80.21	80.65	73.22	69.01	72.12

表 3 误差平方和比较

Table 3 Comparison of error sum of square

数据集	聚类误差平方和			
	PCA-AKM	KNE-KM	QMC-KM	CFSFDP-KM
yeast	0.1031	0.1402	0.1405	0.2714
abalone	0.2275	0.3997	0.4557	0.5122
magic	63.5169	68.1123	69.7319	78.2141
skin	85.1318	91.0412	91.1905	85.1318

由表 2 可得,基于 PCA 的加速 K-means 算法的聚类结果稳定;对于 abalone 数据集,PCA-AKM 算法的聚类精度为 82.13%,达到了 QMC-KM 算法的最高聚类精度,并高于其他两种算法;在 yeast,magic,skin 数据集上,PCA-AKM 的聚类准确率高于其他 3 种算法。

由表 3 可得,PCA-AKM 在 skin 数据集上的误差与 CFSFPD-KM 算法相等,优于 KNE-EM 及 QMC-KM 算法。在 yeast,magic,skin 数据集上,PCA-AKM 算法略优于其他 3 种算法。

5 基于 PCA 的加速 K-means 算法在入侵检测中的应用

本实验采用 KDD CUP 1999[15]数据集中的"kddcup. data_10_percent"作为测试数据,该数据集的攻击类型分为:拒绝服务攻击 DOS、非授权访问 R2L、普通用户非法获得 root 权限 U2R、探测系统漏洞 Probe。该数据集共有 494020 条记录,正常记录数为 97277,其余为入侵记录。每条记录包含 41 个属性项和 1 个类别项。数据特征包括基础特征、网络特征、内容特征[16]。

本实验采用的硬件环境为 3.0GHzCPU,4GB 内存的个人计算机,软件平台为 Windows 7 旗舰版,数据库为 Oracle 11g。实验在 Eclipse 编译环境下实现程序设计。为了评估本算法在入侵检测上的性能,将其分别与传统 K-means 算法、傅涛等人[5]提出的基于 PSO 的 K-means 算法(PSO-based K-means)、于海涛等人[17]提出的基于人工鱼群的优化 K-means聚类算法(AFS-KM)、肖立中等人[18]提出的基于改进粒子群的加速 K 均值算法(NPSO-AKM)进行对比。

采用检测率和误检率两个指标来衡量入侵检测的效果。检测率等于检测到的攻击数目除以所有攻击的总数,误检率等于误报记录数目除以正常记录总数。检测率越高,误报率越低,则代表聚类效果越好,反之亦然。本次实验共执行 30 次,取均值作为最终的检测结果,检测结果如图 1 所示。

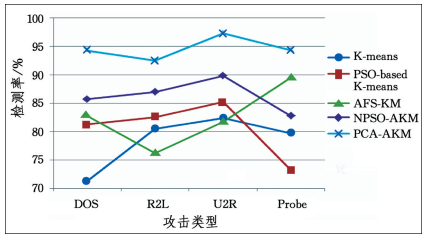

图 1　检测率对比图

Fig. 1　Comparison of detection rate

由图 1 可知，与其他改进的 K-means 算法相比，基于 PCA 的加速 K-means 算法具有较好的检测率，其最大检测率可达到 97.32%，与其他 4 种传统的 K-means 算法相比，检测率分别提高 12.93%，11.97%，14.91%，10.62%。PCA-AKM 针对不同类型的攻击，检测效果也优于其他集中针对传统 K-means 的改进算法。在检测入侵攻击时，PCA-AKM 的检测率较稳定，不会因为攻击类型的不同而使其检测率出现大幅度变化。

5 种 K-means 算法在入侵检测实验中的误检率如图 2 所示，由于传统的 K-means 算法随机选取聚类中心，导致聚类结果不稳定，误检率较高。PSO-based K-means 和 NPSO-AKM 以粒子群方式优化聚类中心，而 AFS-KM 以鱼群方式优化聚类中心，二者都使得误检率降低。本文提出的 PCA-AKM 以密权值指标寻找聚类中心，能够有效防止聚类结果局部最优，最低误检率达到 1.03%，与其他 4 种传统的 K-means 算法相比，误检率分别降低了 7.09%，7.39%，5.78%，6.97%。由此说明，本文提出的基于 PCA 的加速 K-means 算法克服了 K-means 算法随机选取聚类中心而导致的局部最优解问题，聚类效果优于 K-means 算法。

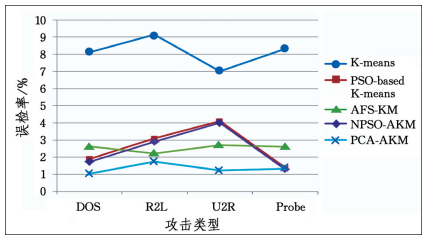

图 2　误检率对比图

Fig. 2　Comparison of error detection rate

K-means 算法的时间复杂度为 $O(nkT)$，n 代表数据样本个数，k 代表簇个数，T 代表算法迭代次数。本文提出的基于 PCA 的 K-means 算法分为两个部分，分别是利用 PCA 进行样本数据降维以及使用密权值选取初始聚类中心。其中，使用 PCA 算法进行数据降维的时间复杂度为 $O(n)$；利用密权值选择初始聚类中心时，需要进行密权值的快速排序，故该部分的时间复杂度为 $O(nlbn)$。算法的整体时间复杂度为 $O(n(1+lbn))$，空间复杂度为 $O(n)$。时间性能比较结果如表 4 所列。可以看出，PCA-AKM 算法的平均检测率与 AFS-KM 算法相近，但执行时间和平均误检率优于 AFS-KM 算法。与其他 3 种算法相比，PCA-AKM 算法在时间性能和检测效果上更优。PCA-AKM 采用主成分分析提取检测数据中的主要成分，再利用密权值选取初始聚类中心进行聚类，不仅提高了检测效率，还提高了检测精度。

表 4　时间性能比较结果

Table 4　Comparison of time performance

算法	执行时间/s	平均检测率/%	平均误检率/%
K-means	273.3	79.24	8.23
NPSO-AKM	251.7	87.14	2.53
PCA-AKM	198.2	93.31	1.30
AFS-KM	360.1	93.27	4.75
PSO-based K-mean	215.3	82.34	7.28

结束语　本文针对传统聚类算法执行效率低的问题，提出使用主成分分析来预先对数据集进行处理的改进 K-means 算法。针对 K-means 算法由随机选择聚类中心造成的聚类结果不稳定问题，本文提出了密权值概念，利用密权值能选择出全局最优的初始聚类中心，以保证聚类的稳定性。将主成分分析与由密权值选取初始聚类中心的 K-means 相结合，提出 PCA-AKM 算法。先利用 Iris 数据集，对本文算法与 K-means在聚类准确率、误差方面进行比较，结果表明本文算法的聚类结果稳定，准确率高于 K-means 算法，且误差较小。在 UCI 数据集上，与 KNE-KM，QMC-KM，CFSFDP-KM 算法进行聚类精度比较，结果表明 PCA-AKM 有更优的聚类效果。采用 KDD CUP99 数据集模拟本文算法在入侵检测中的应用，并与其他 4 种 K-means 改进算法进行比较。实验证明，本文算法在检测率和误检率方面优于其他算法，能取得较好的检测效果。为验证 PCA-AKM 算法的时间性能，与 K-means，NPSO-AKM，AFS-KM，PSO-based K-means 算法进行时间性能对比，结果表明，本文算法在提高执行效率的同时，仍有较优的检测率和误检率，但对于单一属性的入侵数据检测效率仍有待提高。

参 考 文 献

[1] ZHOU W B，SHI Y X. Optimization algorithm of K-means clustering center of selection based on density[J]. Application Research of Computers，2012，29(5)：1726-1728. (in Chinese)
周炜奔，石跃祥. 基于密度 K-means 聚类中心选取的优化算法[J]. 计算机应用研究，2012，29(5)：1726-1728.

[2] ZHAI D H，YU J，GAO F，et al. K-means text clustering algorithm based on initial cluster centers selection according to maximum distance[J]. Application Research of Computers，2014，31(3)：713-715，719. (in Chinese)
翟东海，鱼江，高飞，等. 最大距离法选取初始簇中心的 K-means 文本聚类算法的研究[J]. 计算机应用研究，2014，31(3)：713-715，719.

[3] FENG B，HAO W N，CHEN G，et al. Optimization to K-means initial cluster centers[J]. Computer Engineering and Applications，2013，49(14)：182-185，192. (in Chinese)
冯波，郝文宁，陈刚，等. K-means 算法初始聚类中心选择的优化[J]. 计算机工程与应用，2013，49(14)：182-185，192.

[4] AN J Y，YAN Z J，ZHAI J X. K-means Clustering Algorithm Based on Distance Threshold and Weighted Sample[J]. Microelectronics & Computer，2015(8)：135-138. (in Chinese)
安计勇，闫子骥，翟靖轩. 基于距离阈值及样本加权的 K-means 聚类算法[J]. 微电子学与计算机，2015(8)：135-138.

[5] FU T，SUN Y M. PSO-based k-means Algorithm and its Application in Network Intrusion Detection System[J]. Computer Science，2011，38(5)：54-55，73. (in Chinese)
傅涛，孙亚民. 基于 PSO 的 k-means 算法及其在网络入侵检测

中的应用[J]. 计算机科学,2011,38(5):54-55,73.

[6] YUE S H,WANG J S,TAO G,et al. An unsupervised grid-based approach for clustering analysis[J]. Science China(Information Sciences),2010,53(7):1345-1357.

[7] LI T T. The Research of K-means Clustering Algorithm- Improvement[D]. Hefei:Anhui University,2015. (in Chinese)
李婷婷. 改进 K-means 聚类算法的研究[D]. 合肥:安徽大学,2015.

[8] ZHANG X F,ZHANG G Z,LIU P. Improved K-means algorithm based on clustering criterion function[J]. Computer Engineering and Applications,2011,47(11):123-127. (in Chinese)
张雪凤,张桂珍,刘鹏. 基于聚类准则函数的改进 K-means 算法[J]. 计算机工程与应用,2011,47(11):123-127.

[9] KATSAVOUNIDIS I,JAY KUO C C,ZHANG Z. A new initialization technique for generalized Lloyd iteration[J]. Signal Processing Letters,1994,1(10):144-146.

[10] BOUTSIDIS C,ZOUZIAS A,MAHONEY M W,et al. Randomized Dimensionality Reduction for k-Means Clustering[J]. IEEE Transactions on Information Theory,2011,61(2):1045-1062.

[11] WANG J,KE Q,et al. Fast approximate k-means via cluster closures[C]// IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society,2012:3037-3044.

[12] XIONG K L,PENG J J,YANG X F,et al. K-means Clustering Optimization Based on Kernel Density Estimation[J]. Computer Technology and Development,2017,27(2):1-5. (in Chinese)
熊开玲,彭俊杰,杨晓飞,等. 基于核密度估计的 K-means 聚类优化[J]. 计算机技术与发展,2017,27(2):1-5.

[13] ZHUANG R G,NI Z B,LIU X Y. A Novel Method for Refining the Initial Points for K-means Clustering Based on Quasi-Monte Carlo Method[J]. Journal of University of Jinan(Science and Technology),2017,31(1):35-41. (in Chinese)
庄瑞格,倪泽邦,刘学艺. 基于拟蒙特卡洛的 K 均值聚类中心初始化方法[J]. 济南大学学报(自然科学版),2017,31(1):35-41.

[14] LI M,ZHANG G Z. K-means Algorithm of Optimized Initial Center By Density Peaks[J]. Computer Applications and Software,2017,34(3):212-217. (in Chinese)
李敏,张桂珠. 密度峰值优化初始中心的 K-means 算法[J]. 计算机应用与软件,2017,34(3):212-217.

[15] YUAN T F. Research on Intrusion Detection Based on Data Mining[D]. Chengdu: University of Electronic Science and Technology of China,2014. (in Chinese)
袁腾飞. 基于数据挖掘的入侵检测系统研究[D]. 成都:电子科技大学,2014.

[16] CHENG J. The Research of Fusion Algorithms for Support Vector Machine and K-means Clustering[D]. Dalian: Liaoning Normal University,2008. (in Chinese)
程佳. 支持向量机与 K-均值聚类融合算法研究[J]. 大连:辽宁师范大学,2008.

[17] YU H T,JIA M J,WANG H Q,et al. K-means Clustering Algorithm Based on Artificial Fish Swarm[J]. Computer Science,2012,39(12):60-64. (in Chinese)
于海涛,贾美娟,王慧强,等. 基于人工鱼群的优化 K-means 聚类算法[J]. 计算机科学,2012,39(12):60-64.

[18] XIAO L Z,LIU Y X,CHEN L Q. Research of Accelerating K-Means Algorithm Based on New Particle Swarm Optimization for Intrusion Detection[J]. Journal of System Simulation,2014,26(8):1652-1657. (in Chinese)
肖立中,刘云翔,陈丽琼. 基于改进粒子群的加速 K 均值算法在入侵检测中的研究[J]. 系统仿真学报,2014,26(8):1652-1657.

(上接第 225 页)

MIBS 算法的 6 轮区分器,评估了 MIBS 算法在碰撞攻击下的安全性。分析结果表明,8/9/10 轮的 MIBS 算法对碰撞攻击是不免疫的。

参 考 文 献

[1] IZADI M,SADEGHIYAN B,SADEGHIAN S S,et al. MIBS:a new lightweight block cipher[C]// Proceedings of CANS 2009,Lecture Notes in Computer Science 5888. Berlin:Springer,2009:334-345.

[2] SEBASTIANI F. Machine learning in automated text categorization acmcs[J]. ACM Computing SURCEYS,2002,34(1):1-47.

[3] ZHAO X J,WANG T,WANG S Z,et al. Research on deep differential fault analysis against MIBS[J]. Journal on Communications,2010,31(12):82-88. (in Chinese)
赵新杰,王韬,王素珍,等. MIBS 深度差分故障分析[J]. 通信学报,2010,31(12):82-88.

[4] WANG G L,WANG S H. Integral cryptanalysis of reduced-round MIBS block cipher[J]. Journal of Chinese Computer Systems,2012,33(4):773-777. (in Chinese)
王高丽,王少辉. 对 MIBS 算法的 Integral 攻击[J]. 小型微型计算机系统,2012,33(4):773-777.

[5] LIU C,LIAO F C,WEI H R. Meet-in-the-middle attacks on MIBS[J]. Journal of Inner Mongolia University(Natural Science Edition),2013,44(3):308-313. (in Chinese)
刘超,廖福成,卫宏儒. 对 MIBS 算法的中间相遇攻[J]. 内蒙古大学学报(自然科学版),2013,44(3):308-313.

[6] YU X L,WU W L,LI Y J. Integral cryptanalysis of reduced-round MIBS block cipher[J]. Journal of Computer Research and Development,2013,50(10):2117-2125. (in Chinese)
于晓丽,吴文玲,李俊艳. 低轮 MIBS 分组密码的积分分析[J]. 计算机研究与发展,2013,50(10):2117-2125.

[7] PAN Z S,GUO J S,CAO J K,et al. Integral attack on MIBS block cipher[J]. Journal on Communications,2014,35(7):157-171. (in Chinese)
潘志舒,郭建胜,曹进克,等. MIBS 算法的积分攻击[J]. 通信学报,2014,35(7):157-171.

[8] GILBERT H,MINIER M. A collision attack on 7 rounds of Rijndael[EB/OL]. [2012-10-10]. http://csrc. nist. gov/archive/aes/round2/conf3/papers/11-hgilbert. pdf.

[9] WU W L,FENG D G. Collision attack on reduced-round Camellia[J]. Science in China:Series F,2004,48(1):78-90.

[10] HAN J,ZHANG W J,XU X H. Collision Square Attacks on the Reduced-Round CLEFIA[J]. Acta Electronica Sinica,2009,37(10):2309-2313.

[11] LI C,SUN B,LI R L. Attack method and example analysis of block cipher[M]. Beijing:Science Press,2010:196-199. (in Chinese)
李超,孙兵,李瑞林. 分组密码的攻击方法与实例分析[M]. 北京:科技出版社,2010:196-199.