奇数变元代数免疫最优布尔函数的构造方法

汤 阳 张 宏 张 琨 李千目
(南京理工大学计算机科学与技术学院 南京 210094)

摘 要 代数免疫是随着代数攻击的出现而提出来的一个新的密码学特性。为了有效地抵抗代数攻击,密码系统中使用的布尔函数必须具有最佳的代数免疫。提出了递归构造奇数变元代数免疫最优布尔函数的一个方法。这是一个递归构造的方法,利用该方法,对任意的奇数,都可以构造相同变元数量的代数免疫最优布尔函数。

关键词 密码学,流密码,代数攻击,布尔函数
中图法分类号 TP309.7 **文献标识码** A

Construction of Odd-variable Boolean Function with Optimum Algebraic Immunity

TANG Yang ZHANG Hong ZHANG Kun LI Qian-mu
(School of Computer Science and Technology, Nanjing University of Science and Technology, Nanjing 210094, China)

Abstract Algebraic immunity is a new cryptographic criterion proposed to against algebraic attacks. In order to resist algebraic attacks, Boolean functions used in many stream ciphers should have optimum algebraic immunity. This paper presented a construction of Boolean function in odd variables with optimum algebraic immunity. It's a recursive construction. Given any odd number, we can construct Boolean function with optimum algebraic immunity in the same number of variables.

Keywords Cryptography, Stream cipher, Algebraic attacks, Boolean function

1 引言

从20世纪80年代中期到90年代初期,人们在序列密码的设计方法、分析方法和安全性度量等方面取得了很大进展。布尔函数作为序列密码系统中的一个重要非线性组件,它的性质直接关系到整个密码系统的安全性。为了提高系统的安全强度,布尔函数必须满足一定的准则,如:高代数次数、高非线性度等。对这些性质的研究,均有一些比较成熟的结果[1]。近年来,随着代数攻击的出现,对布尔函数提出新的设计准则和要求——高代数免疫度。为了有效地抵抗代数攻击[2-8],布尔函数必须具有较高的代数免疫度。但是,对于任意一个 n 元布尔函数来说,其代数免疫度有上界 $\lceil n/2 \rceil$[5]。布尔函数的代数免疫度达到上界 $\lceil n/2 \rceil$ 时,就称为代数免疫最优的[5]。显然,代数免疫最优的函数具有最佳的抵抗代数攻击的能力。因此,如何构造代数免疫最优的布尔函数[9-14]是代数攻击领域的一个研究热点。

本文提出了一个构造奇数变元代数免疫最优布尔函数的方法。根据该方法,可以构造任意奇数变元的代数免疫最优布尔函数。本文首先介绍布尔函数的基本知识;然后提出一类递归构造奇数变元布尔函数的方法,并对函数的代数免疫最优性进行证明;最后总结全文。

2 预备知识

定义1 设 n 为正整数,$GF(2)$ 是有限域。n 元布尔函数定义为如下的映射:

$$f: GF(2)^n \rightarrow GF(2)$$

记为 $f(x)$,其中 $x \in GF(2)^n$,$f(x) \in GF(2)$。记 B_n 为所有 n 元布尔函数的集合。

给定 $GF(2)^n$ 上的一个序,对于一个 n 元的布尔函数 $f(x)$,按 $GF(2)^n$ 上向量从小到大的顺序,可将其函数值从左到右排列合成一个向量 T。向量 T 就称为布尔函数 $f(x)$ 的真值表。显然,T 是一段长度为 2^n 的01串。

定义2 任意一个 n 元布尔函数 $f(x)$,都可唯一表示为 $GF(2)$ 上一个次数不超过 n 的多项式

$$f(x)=a_0+\sum_{1\leqslant i\leqslant n}a_i x_i+\sum_{1\leqslant i<j\leqslant n}a_{i,j}x_i x_j+\cdots+a_{1,2,\cdots,n}x_1 x_2\cdots x_n \quad (1)$$

式中,$a_0, a_i, a_{i,j}, \cdots, a_{1,2,\cdots,n} \in GF(2)$。这种表示方式称为 $f(x)$ 的代数正规型表示。式(1)中系数不为零的项的最大次数称为 $f(x)$ 的代数次数,记为 $\deg(f)$。

定义3[10] n 元布尔函数 f 的零化集 $AN(f)$ 定义为

$$AN(f)=\{g \in B_n \mid g \cdot f=0\}$$

并且对任意的 $g \in AN(f)$,称 g 是 f 的零化子。

到稿日期:2010-04-19 返修日期:2010-07-14 本文受国家自然科学基金(60903027)资助。

汤 阳(1977—),男,博士生,主要研究方向为信息安全,E-mail:yang.t@sohu.com;张 宏(1956—),男,教授,博士生导师,主要研究方向为信息安全、数据挖掘;张 琨(1977—),女,博士,副教授,主要研究方向为信息安全、自主计算;李千目(1979—),男,博士,副教授,主要研究方向为信息安全、信息系统工程。

显然，$g=0$ 是所有布尔函数的零化子，这是一个平凡零化子。

定义 4[10] 设 f 为 n 元布尔函数，其代数免疫(度)定义为

$$\mathrm{AI}(f)=\min\{\deg(g)\mid 0\neq g\in AN(f)\cup AN(f+1)\}$$

文献[2]指出，任意一个 n 元布尔函数都具有一个代数次数不超过$\lceil n/2\rceil$的零化子。所以，$\mathrm{AI}(f)\leqslant\lceil n/2\rceil$。因此，我们把满足 $\mathrm{AI}(f)=\lceil n/2\rceil$的 f 称为代数免疫最优的。

设 s_1 和 s_2 都是长为 2^n 的 01 串，用"$s_1\|\|s_2$"表示 s_1 和 s_2 顺序连接成串。这是长为 2^{n+1}的 01 串，表示了一个 $n+1$ 元的布尔函数。关于记号"$\|\|$"，可任意证明以下性质成立。

性质 1 假设 $f=f_1\|\|f_2\|\|f_3\|\|f_4$，其中 $f_1,f_2,f_3,f_4\in B_n$，那么

i) $f\in B_{n+2}$，且 $f=f_1+x_{n+1}(f_1+f_2)+x_{n+2}(f_1+f_3)+x_{n+1}x_{n+2}(f_1+f_2+f_3+f_4)$；

ii) 如果 $g\in AN(f)$，且记 $g=g_1\|\|g_2\|\|g_3\|\|g_4$，$g_1,g_2,g_3,g_4\in B_n$，则 $g_1\in AN(f_1)$，$g_2\in AN(f_2)$，$g_3\in AN(f_3)$，$g_4\in AN(f_4)$。

3 构造奇数变元代数免疫最优的布尔函数

Dalai 等[9,10]用递归构造的方法，构造了偶数变元代数免疫最优的布尔函数。本文引用其递归构造的思想，给出了如下递归构造奇数变元布尔函数的构造方法。

构造 1：

$$\begin{cases}\phi_{2n-1}=\phi_{2n-1}\|\|\phi_{2n-1}\|\|\phi_{2n-1}\|\|\phi_{2n-1}^1\\ \phi_{2n+1}^i=\phi_{2n-1}^{i-1}\|\|\phi_{2n-1}^i\|\|\phi_{2n-1}^i\|\|\phi_{2n-1}^{i+1}\end{cases}\tag{2}$$

式中，$\phi_n^0=\phi_n$，$\phi_n^j=x_1+1+(j \bmod 2)$，$i,n\geqslant 1$，$j\geqslant 0$。

为了方便起见，我们定义 $\phi_n^{-1}=\phi_n^0=\phi_n$。于是，递归式(2)可以简化为

$$\phi_{2n+1}^i=\phi_{2n-1}^{i-1}\|\|\phi_{2n-1}^i\|\|\phi_{2n-1}^i\|\|\phi_{2n-1}^{i+1}\tag{3}$$

下面，如果没有特别指出，符号 ϕ 均表示构造 1 中的布尔函数。在下面的证明中，可能会碰到一些不可能出现的布尔函数(比如，一个代数次数小于零的布尔函数)，这时用 0 函数来代替这些函数。

引理 1 已知 $i\geqslant 0$，假如对 $\forall 0\leqslant t\leqslant i$ 都有 $AI(\phi_{2t+1})=t+1$，那么对任意的 $g\in AN(\phi_{2i+1}^j)$，$h\in AN(\phi_{2i+1}^{j+1})$，如果满足 $\deg(g+h)\leqslant i-1-j$，$j\geqslant 0$，那么有 $g=h$。

证明：对 i 进行归纳。

当 $i=0$ 时，$\deg(g+h)\leqslant i-1-j=0-1-j\leqslant -1$，不存在这样的 g 和 h，故 $g=0$，$h=0$。

假设当 $i<k$ 时，命题均成立。

那么当 $i=k$ 时，对 $g,h\in B_{2k+1}$，不妨记

$$\begin{cases}g=g_1\|\|g_2\|\|g_3\|\|g_4\\ h=h_1\|\|h_2\|\|h_3\|\|h_4\end{cases}$$

式中，$g_1,g_2,g_3,g_4,h_1,h_2,h_3,h_4\in B_{2k-1}$。

于是根据性质 1，有

$$\begin{aligned}g+h=&(g_1+h_1)+x_{2k}(g_1+h_1+g_2+h_2)+x_{2k+1}(g_1+h_1+\\&g_3+h_3)+x_{2k}x_{2k+1}(g_1+h_1+g_2+h_2+g_3+h_3+\\&g_4+h_4)\end{aligned}\tag{4}$$

假设 $g\in AN(\phi_{2k+1}^j)$，$h\in AN(\phi_{2k+1}^{j+1})$，$\deg(g+h)\leqslant k-1-j$，$j\geqslant 0$。

此时，对 $j\geqslant 0$，根据递归关系(3)和性质 1 有

$$\begin{cases}\phi_{2k+1}^j=\phi_{2k-1}^{j-1}\|\|\phi_{2k-1}^j\|\|\phi_{2k-1}^j\|\|\phi_{2k-1}^{j+1}\\ \phi_{2k+1}^{j+1}=\phi_{2k-1}^j\|\|\phi_{2k-1}^{j+1}\|\|\phi_{2k-1}^{j+1}\|\|\phi_{2k-1}^{j+2}\end{cases}$$

并且

$$\begin{cases}g_1\in AN(\phi_{2k-1}^{j-1}),g_2,g_3\in AN(\phi_{2k-1}^j),g_4\in AN(\phi_{2k-1}^{j+1})\\ h_1\in AN(\phi_{2k-1}^j),h_2,h_3\in AN(\phi_{2k-1}^{j+1}),h_4\in AN(\phi_{2k-1}^{j+2})\end{cases}$$

a) 从式(4)知，

$\deg(g_1+h_1)\leqslant\deg(g+h)\leqslant k-1-j=(k-1)-1-(j-1)$

如果 $j>0$，因为 $g_1\in AN(\phi_{2k-1}^{j-1})$，$h_1\in AN(\phi_{2k-1}^j)$，由归纳假设，可得 $g_1=h_1$。

如果 $j=0$，则 $g_1,h_1\in AN(\phi_{2k-1})$，进而 $g_1+h_1\in AN(\phi_{2k-1})$。

根据命题假设有 $AI(\phi_{2k-1})=k$，而 $\deg(g_1+h_1)\leqslant k-1<AI(\phi_{2k-1})$，所以 $g_1+h_1=0$，即 $g_1=h_1$。

b) 把 $g_1+h_1=0$ 代入式(4)，可得

$$\begin{aligned}g+h=&x_{2k}(g_2+h_2)+x_{2k+1}(g_3+h_3)+x_{2k}x_{2k+1}(g_2+h_2+\\&g_3+h_3+g_4+h_4)\end{aligned}\tag{5}$$

由上式可得，$\deg(g_2+h_2)\leqslant\deg(g+h)-1\leqslant(k-1-j)-1=(k-1)-1-j$。

而 $g_2\in AN(\phi_{2k-1}^j)$，$h_2\in AN(\phi_{2k-1}^{j+1})$，由归纳假设得 $g_2=h_2$。

类似地，有 $\deg(g_3+h_3)\leqslant(k-1)-1-j$，且 $g_3\in AN(\phi_{2k-1}^j)$，$h_3\in AN(\phi_{2k-1}^{j+1})$，由归纳假设得 $g_3=h_3$。

c) 由于 $g_2=h_2$，$g_3=h_3$，式(5)化简为

$$g+h=x_{2k}x_{2k+1}(g_4+h_4)\tag{6}$$

此时，$\deg(g_4+h_4)=\deg(g+h)-2\leqslant(k-1-j)-2=(k-1)-1-(j+1)$。

而 $g_4\in AN(\phi_{2k-1}^{j+1})$，$h_4\in AN(\phi_{2k-1}^{j+2})$，由归纳假设得 $g_4=h_4$。

因此 $g=h$。证毕。

引理 2 已知 $i\geqslant 0$，假如对 $\forall 0\leqslant t\leqslant i$ 都有 $AI(\phi_{2t+1})=t+1$，那么对任意的 $g\in AN(\phi_{2i+1}^j)\cap AN(\phi_{2i+1}^{j+1})$，如果满足 $\deg(g)\leqslant i+j+1$，$j\geqslant 0$，那么有 $g=0$。

证明：对 i 进行归纳。

当 $i=0$ 时，由于 $\phi_1^j=x_1+1+(j \bmod 2)$，因此 $\phi_1^{j+1}=\phi_1^j+1$。

$g\in AN(\phi_1^j)\cap AN(\phi_1^{j+1})=AN(\phi_1^j)\cap AN(\phi_1^j+1)$，因而 $g=0$。

假设当 $i<k$ 时，命题对所有的 $j\geqslant 0$ 均成立。

那么当 $i=k$ 时，对 $g\in B_{2k+1}$，不妨记

$$g=g_1\|\|g_2\|\|g_3\|\|g_4$$

式中，$g_1,g_2,g_3,g_4\in B_{2k-1}$。

根据性质 1，有

$$\begin{aligned}g=&g_1+x_{2k}(g_1+g_2)+x_{2k+1}(g_1+g_3)+x_{2k}x_{2k+1}\\&(g_1+g_2+g_3+g_4)\end{aligned}\tag{7}$$

假设 $g\in AN(\phi_{2k+1}^j)\cap AN(\phi_{2k+1}^{j+1})$，$\deg(g)\leqslant k+j+1$，$j\geqslant 0$。

根据递归关系(3)和性质 1 有

$$\begin{cases}\phi_{2k+1}^j=\phi_{2k-1}^{j-1}\|\|\phi_{2k-1}^j\|\|\phi_{2k-1}^j\|\|\phi_{2k-1}^{j+1}\\ \phi_{2k+1}^{j+1}=\phi_{2k-1}^j\|\|\phi_{2k-1}^{j+1}\|\|\phi_{2k-1}^{j+1}\|\|\phi_{2k-1}^{j+2}\end{cases}$$

且

$$\begin{cases}g_1\in AN(\phi_{2k-1}^{j-1})\cap AN(\phi_{2k-1}^j)\\ g_2,g_3\in AN(\phi_{2k-1}^j)\cap AN(\phi_{2k-1}^{j+1})\\ g_4\in AN(\phi_{2k-1}^{j+1})\cap AN(\phi_{2k-1}^{j+2})\end{cases}$$

a)由式(7)知，

$\deg(g_1)\leqslant\deg(g)$

$\deg(g_1+g_2),\deg(g_1+g_3)\leqslant\deg(g)-1<\deg(g)$

$\deg(g_1+g_2+g_3+g_4)\leqslant\deg(g)-2<\deg(g)$

所以，$\deg(g_2),\deg(g_3)\leqslant\deg(g)$。

进而，$\deg(g_4)\leqslant\deg(g)\leqslant k+j+1=(k-1)+(j+1)+1$。

又 $g_4\in AN(\phi_{2k-1}^{j+1})\cap AN(\phi_{2k-1}^{j+2})$，根据归纳假设，有 $g_4=0$。

b) 把 $g_4=0$ 代入式(7)，得

$$g=g_1+x_{2k}(g_1+g_2)+x_{2k+1}(g_1+g_3)+x_{2k}x_{2k+1}(g_1+g_2+g_3) \tag{8}$$

于是，

$\deg(g_1+g_2),\deg(g_1+g_3)\leqslant\deg(g)-1\leqslant k+j$

$\deg(g_1+g_2+g_3)\leqslant\deg(g)-2\leqslant k+j-1<k+j$

进而有，$\deg(g_1),\deg(g_2),\deg(g_3)\leqslant k+j=(k-1)+j+1$。

又 $g_2\in AN(\phi_{2k-1}^{j})\cap AN(\phi_{2k-1}^{j+1})$，根据归纳假设，$g_2=0$。

同理，有 $g_3=0$。

c) 把 $g_2=g_3=0$ 代入式(8)，得

$$g=g_1(1+x_{2k}+x_{2k+1}+x_{2k}x_{2k+1}) \tag{9}$$

所以，$\deg(g_1)=\deg(g)-2\leqslant k+j-1$。

当 $j>0$ 时，$g_1\in AN(\phi_{2k-1}^{j-1})\cap AN(\phi_{2k-1}^{j})$。

由于 $\deg(g_1)\leqslant k+j-1=(k-1)+(j-1)+1$，根据归纳假设，有 $g_1=0$。

当 $j=0$ 时，$g_1\in AN(\phi_{2k-1})$。

根据命题假设有 $AI(\phi_{2k-1})=k$，而 $\deg(g_1)\leqslant k-1<AI(\phi_{2k-1})$，所以 $g_1=0$。

把 $g_1=0$ 代入式(9)，可得 $g=0$。

证毕。

定理 1 布尔函数 $\phi_{2k+1}\in B_{2k+1}$，$k\geqslant 0$ 代数免疫最优，即 $AI(\phi_{2k+1})=k+1$。

证明：对 k 进行归纳。

当 $i=0$ 时，容易验证命题成立。

假设原命题在 $i<k$ 时均成立。

那么当 $i=k$ 时，

对任意的 $g\in AN(\phi_{2k+1})$，如果 $\deg(g)\leqslant k$，我们证明 $g=0$。

不妨记

$$g=g_1||g_2||g_3||g_4$$

式中，$g_1,g_2,g_3,g_4\in B_{2k-1}$。

由于 $\phi_{2k+1}=\phi_{2k-1}||\phi_{2k-1}||\phi_{2k-1}||\phi_{2k-1}^{1}$，根据性质 1 有

$g_1,g_2\in AN(\phi_{2k-1}),g_3\in AN(\phi_{2k-1}),g_4\in AN(\phi_{2k-1}^{1})$

且

$$g(x_1,x_2,\cdots,x_{2k},x_{2k+1})=g_1+x_{2k}(g_1+g_2)+x_{2k+1}(g_1+g_3)+x_{2k}x_{2k+1}(g_1+g_2+g_3+g_4) \tag{10}$$

由于 $\deg(g)\leqslant k$，因此

$\deg(g_1+g_2),\deg(g_1+g_3)\leqslant\deg(g)-1\leqslant k-1$

$\deg(g_1+g_2+g_3+g_4)\leqslant\deg(g)-2\leqslant k-2$

从 $g_1,g_2\in AN(\phi_{2k-1})$ 知 $g_1+g_2\in AN(\phi_{2k-1})$。

根据命题假设有 $AI(\phi_{2k-1})=k$，而 $\deg(g_1+g_2)\leqslant k-1<AI(\phi_{2k-1})$，故 $g_1+g_2=0$。

所以式(10)化简成

$$g=g_1+x_{2k}(g_1+g_3)+x_{2k}x_{2k+1}(g_3+g_4) \tag{11}$$

于是，$\deg(g_3+g_4)\leqslant\deg(g)-2\leqslant k-2=(k-1)-1$。

考虑到 $g_3\in AN(\phi_{2k-1}),g_4\in AN(\phi_{2k-1}^{1})$，根据引理 1，有 $g_3=g_4$。

于是，$g_3\in AN(\phi_{2k-1})\cap AN(\phi_{2k-1}^{1})$。

从式(11)知

$\deg(g_1)\leqslant\deg(g)\leqslant k,\deg(g_1+g_3)\leqslant\deg(g)-1\leqslant k-1<k$

进而 $\deg(g_3)\leqslant k=(k-1)+1$。根据引理 2，便有 $g_3=0$。

把 $g_3=g_4=0$ 代入式(11)得

$$g=g_1(1+x_{2k}) \tag{12}$$

于是，$\deg(g_1)=\deg(g)-1\leqslant k-1$。

而 $g_1\in AN(\phi_{2k-1})$，且 $AI(\phi_{2k-1})=k>\deg(g_1)$，所以 $g_1=0$，从而 $g=0$。

所以，对任意的 $0\neq g\in AN(\phi_{2k+1})$，都有 $\deg(g)\geqslant k+1$。

同理可证，对任意的 $0\neq g\in AN(\phi_{2k+1}+1)$，也都有 $\deg(g)\geqslant k+1$。

因此，$AI(\phi_{2k+1})=k+1$。证毕。

结束语 本文提出了构造奇数变元代数免疫最优布尔函数的方法。这是一个递归构造的方法，利用该方法，对任意的奇数，都可以构造相应变元数量的代数免疫最优的布尔函数。这在代数攻击的背景下，具有相当的实际意义。

参 考 文 献

[1] 温巧燕，钮心忻，杨义先. 现代密码学中的布尔函数[M]. 北京：北京科学出版社，2000

[2] Courtois N, Meier W. Algebraic Attacks on Stream Ciphers with Linear Feedback [C] // Advances in Cryptology-Eurocrypt, 2003. Berlin: Springer-Verlag, 2003: 345-359

[3] Courtois N, Klimov A, Patarin J, et al. Efficient algorithms for solving overdefined systems of multivariate polynomial equations [C]// Advances in Cryptology-Eurocrypt, 2000. Berlin: Springer-Verlag, 2000: 392-407

[4] Kipnis A, Shamir A. Cryptanalysis of the HFE public key cryptosystem by relinearization[C] // Advances in Cryptology-Crypto'99. Berlin: Springer-Verlag, 1999: 19-30

[5] Meier W, Pasalic E, Carlet C. Algebraic attacks and decomposition of Boolean functions[C] // Advances in Cryptology-Eurocrypt, 2004. Berlin: Springer-Verlag, 2004: 474-491

[6] Armknecht F, Krause M. Algebraic Attacks on Combiners with Memory[C] // Advances in Cryptology-Crypto, 2003. Berlin: Springer-Verlag, 2003: 162-175

[7] Courtois N. Algebraic Attacks on Combiners with Memory and Several Outputs [C] // Information security and cryptology-ICISC 2004. 2005Berlin: Springer-Verlag, 2005: 3-20

[8] Armknecht F. On the Existence of low-degree Equations for Algebraic Attacks[EB/OL]. http://eprint.iacr.org/2004/185

[9] Dalai D K, Gupta K C, Maitra S. Cryptographically Significant Boolean functions: Construction and Analysis in terms of Algebraic Immunity[C] // Fast Software Encryption. Berlin: Springer-Verlag, 2005: 98-111

[10] Carlet C, Dalai D K, Gupta K C, et al. Algebraic Immunity for Cryptographically Significant Boolean Functions: Analysis and Construction[J]. IEEE Transactions on Information Theory, 2006, 52(7): 3105-3121

[11] Chen Y D, Lu P Z. Constructions of Even-variable Boolean Function with Optimum Algebraic Immunity[EB/OL]. http://eprint.iacr.org/2009/130

[12] Bracken A, Preneel B. On the algebraic immunity of symmetric Boolean functions[C]//Progress in Cryptology-Indocrypt 2005. Berlin: Springer-Verlag, 2005: 35-48

[13] Dalai D K, Maitra S, Sarkar S. Basic Theory in Construction of Boolean Functions with Maximum Possible Annihilator Immunity[J]. Design, Codes and Cryptography, 2006, 40(1): 41-58

[14] Qu L, Li C. On the 2^m-variable symmetric Boolean functions with maximum algebraic immunity[J]. Science in China Information Sciences, 2008, 51(2): 120-127

(上接第 82 页)

NFCD 算法的整体准确率如表 2 所列，对于该算法，聚类结果的数目和聚类参数是自动确定的。从表 2 中的实验结果可以看出，相对于 K-Means 和 DBSCAN 算法，NFCD 具有最高的准确率。对 zsdx09 和 auck8 数据集，算法的平均准确率分别到达 94.6%和 95.7%。

表 2 NFCD 算法的整体准确率(OA)

Data set	Average	Minimum	Maximum
auck8	95.7%	93.2%	97.6%
zsdx09	94.6%	91.3%	98.0%

图 3 给出了使用数据集 zsdx09 时，DBSCAN，K-Means 和 NFCD 算法对各个应用的精确率。对于这几种应用，NFCD 算法在各应用类别中都取得了最佳的结果。

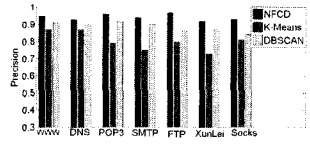

图 3 DBSCAN，K-Means 和 NFCD 的精确率

4.3.3 聚类效果

在处理流量聚类过程中，聚类算法所产生的聚类数量应加以重点考察。因为当聚类完成后，每一个类都要进行标记，聚类的数量越小越容易标记，所以减少聚类数量非常关键。图 4 使用 zsdx09 数据集，表示了流数目的百分比和聚类数目的百分比的关系。在这次聚类过程中 K-Means 算法的输入参数 k 设置为 100。DBSCAN 算法的输入参数设置为 eps=0.03，minPts=12。从图可知，NFCD 相对于 K-Means 和 DBSCAN 算法，可产生更好的聚集效果。对于 NFCD 算法，最大的 6 个社团包含了接近 70%的流。这 6 个社团分别是 HTTP，DNS，PoP3，SMTP，FTP，Socks，其整体准确率也高达 96%。对 auck8 数据集进行类似实验，得到相似的结果。

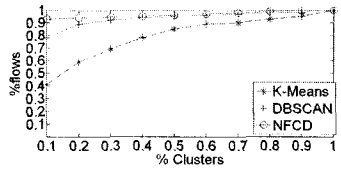

图 4 聚类效果的 CDF

另一个显著的不同就是算法的聚类时间，若使用 auck8 数据集子集进行测试，K-Means 算法是最快的，平均时间不到 56 秒，NFCD 算法用时也较短，平均时间为 72 秒，DBSCAN 算法的时间相对较长，平均为 150 秒。

结束语 网络流量分类与应用识别对于网络管理、安全、研究等非常重要。简单的基于端口号分类的方法逐渐失去效用，基于深度分组的分类技术也存在各种缺点。本文根据不同类别应用产生的网络流量具有独特的统计特征的性质，提出了一种基于复杂网络社团划分算法的无监督的网络流量分类方法，并与先前应用于流量分类的 K-Means 和 DBSCAN 算法进行了比较。通过实际数据集验证了本文提出的方法在准确性和聚类效果上具有明显的优越性。

参 考 文 献

[1] Sen S, Spatscheck O, Wang Dongmei. Accurate, scalable in network identification of p2p traffic using application signature[C]// Proceedings of the 13th international conference on World Wide Web. 2004: 512-521

[2] Haffner P, Sen S, Spatscheck O, et al. ACAS: Automated Construction of Application signatures[C]//Proceedings of the 2005 ACM SIGCOMM Workshop on Mining Network Data. 2005: 197-202

[3] Moore A, Zuev D. Internet traffic classification using Bayesian analysis techniques [C] // ACM International Conference on Measurement and Modeling of Computer Systems (SIGMERTRICS). 2005: 50-60

[4] 徐鹏，林森. 基于 C4.5 决策树的流量分类方法[J]. 软件学报，2009(10): 2692-2704

[5] Williams N, Zander S, Armitage G. A preliminary performance comparison of five machine learning algorithmsfor practical IP traffic flow classification[J]. Special Interest Group on Data Communication Computer Communication Review, 2006: 5-15

[6] Auld T, Moore A W, Gull S F. Bayesian neural networks for Internet traffic classification[J]. IEEE Transactions on Neural Networks, 2007: 223-239

[7] Erman J, Alitt M, Mahanti A. Traffic classification using clustering algorithms[C]// ACM SIGCOMM MineNet. 2006: 281-286

[8] Strehl A, Ghosh J, Mooney R. Impact of similarity measures on web-page clustering [D]. AI for Web Search, 2000

[9] Newman M E J. Mixing patterns in networks[J]. Phys. Rev. E, 2003, 67: 026126

[10] Clauset A, Newman M E J, Moore C. Finding community structure in very large networks[J]. Phys. Rev. E, 2004, 70: 066111

[11] NLANR-Passive Measurement and Analysis[OL]. http://pma.nlanr.net

[12] Iffeit J. Principal component analysis (2^{nd}) [M]. New York: Springer-Verlag, 2003

[13] Zander S, Nguyen T, Armitage G. Automated Traffic Classification and Application Identification Using Machine Learning [C]// The IEEE Conference on Local Computer Networks 30th Anniversary. 2005: 250-257

[14] Moore A W, Papagiannaki K. Toward the accurate identification of network applications[C]//Dovrolis C, ed. Proc. of the PAM 2005. LNCS 3431. Heidelberg: Springer-Verlag, 2005: 41-54

[15] Ester M, Kriege H P, Sander J. A density-based algorithm for discovering clusters in large spatial databases with noise[C]// Proc. of the 2^{nd} International Confabulation Knowledge Discovery and Data Mining Portland. 1996: 226-231