有组织犯罪集团挖掘算法①

唐德权[1], 史伟奇[1], 凌志刚[2]

[1](湖南警察学院 信息技术(网监)系, 长沙 410138)
[2](湖南大学 电气与信息工程学院, 长沙 410082)

摘 要: 互联网的发展使有组织犯罪集团的认定与处罚变的更为困难. 针对有组织犯罪集团的特点, 采用共犯网络结构分析和数据挖掘技术相结合的方法确定有组织犯罪集团结构及其组成实体, 提出一种对有组织犯罪集团进行系统分析和挖掘的新算法. 该算法能从大型真实犯罪数据集提取信息, 快速获取有组织犯罪集团证据, 提高了有组织犯罪集团检测效率. 最后将算法与其他现有算法进行比较, 实验结果表明该算法的时间性能优越.

关键词: 有组织犯罪集团; 共犯矩阵; 挖掘技术; 犯罪证据; 时间性能

Algorithm for Ming Organized Crime

TANG De-Quan[1], SHI Wei-Qi[1], LING Zhi-Gang[2]

[1](Department of Information Technology, Hunan Police Academy, Changsha 410138, China)
[2](College of Electrical and Information Engineering, Hunan University, Changsha 410082, China)

Abstract: It is more difficult that cognizance and punishment of organized crime groups with development of the Internet. According to the characteristics of the organized crime groups, we are using co-offending networks analysis methods and data mining techniques to identifying organized crime structures and their constituent entities. An novel algorithm for mining organized crime groups is proposed. The goal of our work is to improve the efficiency of the organized crime detection for extracting information from large real-life crime datasets to obtain evidence of the organized crime group. Experimental results show that the algorithm of time performance is superior compared with other existing algorithms.

Key words: organized crime groups; co-offending matrix; mining technology; crime evidence; time performance

1 引言

共犯网络作为有组织犯罪集团分析的重要手段, 已经成为当前热门研究领域[1]. 由于共犯网络也构成一个广泛的社会网络, 近年来在反恐、国家公共安全、军事等方面具有广阔的应用前景. 目前, 国内外对共犯网络的学术研究和社会意识已经提高, 执法部门和情报机构早已意识到共犯网络分析的潜力, 经过近年的发展, 它逐渐成为了研究有组织犯罪集团证据的主流, 能提供一种以证据为基础打击和预防犯罪有效方法. Morselli[2]从犯罪网络的角度, 应用共犯网络分析了大量犯罪团体和组织案例, 对犯罪组织进行了深入的研究并提出了"犯罪组织系统"见解, 但是实验数据只对 205 个样本集进行, 并不代表一般犯罪团体特别是大型犯罪集团. Reiss[3]的结论是大多数共犯组织具有不稳定性, 他们的关系是短暂的, 这证实了一些频繁罪犯的共犯随着时间的推移关系稳定, 但并未确定共犯有许多不同的合作同伙, 随着时间的推移不太可能与同一个人进行犯罪行为.

近年来, 我国对有组织犯罪心理、行为和相关数据的规律也进行了大量研究工作. 马万等提出社会网络结构对有组织犯罪集团调查取证十分重要[4]; 温粉莲、唐常杰等提出一种基于最短路径算法 SPLINE

① 基金项目:国家高技术研究发展计划(863)(2012AA112312);教育部高等学校博士学科点专项科研基金(20110161120006);湖南省公安厅科研基金(湘公科信明电(2013)56 号);湖南省教育规划课题(XJK013CXX012)
收稿时间:2015-01-28;收到修改稿时间:2015-03-18

(Shortest Path algorithm based on LINk wEight), 利用 Dijistra 算法计算集团网络任意两个结点间的最短路径, 小于阈值的最短路径保留作为犯罪集团数据挖掘的结果[5]; 刘齐宏等在 SPLINE 算法基础上通过中心度衡量找到核心成员, 提出了挖掘犯罪集团核心成员算法 CMM (Core Members Mining algorithm) [6]. 这些算法在有效性和预测犯罪核心的准确率方面都有待改进.

因此, 如何从复杂犯罪网络图结构中, 快速计算有组织犯罪集团子结构并准确计算核心成员是有组织犯罪集团挖掘方法的关键. 本文采用犯罪网络分析和数据挖掘技术相结合, 提出一种新的有组织犯罪集团检测计算方法 OCMM(Organized Crime Ming Method). 该方法与传统的犯罪分析方法相比, 能有效的提取大型真实犯罪组织数据集特别是警方报告犯罪数据集的信息, 极大提高了犯罪核心网络结构分析计算效率和准确率.

2 基本知识

本文的犯罪数据模型 $H(N,\varepsilon)$用图论的二元结构形式化表示, 其中 N 为结点集, ε 为超边集. 结点集 N 分成三个子集: $A=\{a_1,a_2, \ldots,a_q\}$, $I=\{i_1,i_2, \ldots,i_r\}$, $R=\{r_1,r_2, \ldots,r_s\}$, 分别代表人物(如罪犯、受害者、目击者、嫌疑人和旁观者等), 犯罪事件(即每个事件的犯罪类型)和犯罪资源(如移动电话、工具、车辆、武器或银行账户等). 超边集合 ε 中 e 是非空结点子集 $\{n_1,n_2, \ldots,n_p\} \subseteq N$ 构成边, 必须满足下列三个条件: $|e\cap I|=1$, $|e\cap A|\geq 1$ 和 $|e\cap R|\geq 1$.

共犯网络一般从犯罪数据模型中产生, 由一个或多个结点组成, 这里结点是将罪犯连接在一起[7]. 共犯的数量通过共犯结点 u,v 链接关联一个强度值 $l=\{u,v\}$表示, 记为 $S(l), S(l)\in N$. 假设有 k 个罪犯和 m 个犯罪事件$(k, m > 1)$, 一个 $k \times m$ 矩阵 M, 如果罪犯 O_u 参与事件 i_v, 则 $m_{uv}=1$,否则为 0. 因此一个共犯网络是 $k \times k$ 矩阵: $N=MM^T$, 如公式(1).

$$n_{u,v} = \sum_{x=1}^{k} n_{ux} n_{xv} \tag{1}$$

定义 1. 犯罪活动

对于某个罪犯组 C_i^t,记录这一组的成员在 t_1 至 t_2 时间犯下了罪行的犯罪频率, 称为 C_i^t 的犯罪活动, 用 $\theta_i^{t_1,t_2}$ 表示.

定义 2. 团体犯罪程度

团体犯罪程度用 Φ_I^t 表示, 为了衡量罪犯组 C_i^t 成员在时间 t 犯下的罪行的严重性程度.

定义 3. 活跃罪犯集团

活跃罪犯集团是指在一定时间范围罪犯集团持续犯罪活动, 用 $A_I^{t_1,t_2}$ 表示一个从时间 t_1 到时间 t_2 持续活动犯罪集团.

定义 4. 严重的犯罪集团

一个犯罪组织的总体犯罪活动在时间 t 表现出高度的严重刑事犯罪被称为严重犯罪集团, 用 S_I^t 表示.

定义 5. 有组织犯罪集团演变

有组织犯罪集团 O_a^t, 从时间 t 至 n 连续时间段动态转换或演变序列为 $O_a^t, O_{a_1}^{t+1}, O_{a_2}^{t+2}, \ldots, O_{a_n}^{t+n}$, 用 $E\left(O_a^t\right)$表示有组织犯罪演变.

3 有组织犯罪集团挖掘

网络社团检测计算是社交网络里一个热点研究课题[8], 有组织犯罪集团的性质不同于其他类型如朋友圈或合作组织, 有组织犯罪集团通常有清晰和严格定义的组织成员, 犯罪集团的成员为实现物质利益有着紧密联系的系统组织. 因此, 检测有组织犯罪集团要求更严格的犯罪集团定义.

根据犯罪学中基本理论, 可以总结有组织的犯罪集团重要特征为: 1)这些组织至少有三名成员, 可以分为集中式或分布式或等级式. 不管怎么分类, 重点是研究犯罪集团之间协作的密度比犯罪集团内部协作的密度较高; 2)犯罪组织中个人角色在不同组织程度的机构之间分布特点不同, 充当角色可以重叠也可能是共同成员; 3)这些团伙组织为获得物质利益都犯下严重罪行; 4)相比普通罪犯组他们的活动更加连续.

3.1 有组织犯罪集团计算

犯罪活动和犯罪行为是理解犯罪集团组织结构两个关键特征[9]. 下面提出两个操作算子对犯罪活动和行为进行计算.

犯罪集团 C_i 在时间 t 犯罪行为表示为 $\Phi(C_i)$, 定义为:

$$\Phi(C_i) = \sum_{k=1}^{k=n} \frac{\phi_k}{n} \tag{2}$$

这里 ϕ_{i_k} 表示某个罪犯 i_k 的严重程度, 即犯罪集团 C_i 成员在 t 时刻的犯罪行为.

设 $i_1,i_2,...,i_n$ 是 C_i 在时间 t 的犯罪成员, 犯罪集团 C_i 在时间 t_1 到时间 t_2 的活动记为 $\theta_{t_1,t_2}(C_i)$, 计算公式如下:

$$\theta_{t_1,t_2}(C_i)=\frac{|R_{t_1}(C_i)|}{|R_{t_2}(C_i)|} \tag{3}$$

这里 $|R_{t_1}(C_i)|$ 和 $|R_{t_2}(C_i)|$ 分别表示犯罪集团 C_i 在时刻 $t1$ 和时刻 $t2$ 共犯次数.

为了确定发现罪犯组是否被认为是有组织犯罪集团, 必须同时考虑犯罪活动和犯罪行为, 定义两个阈值: α 表示犯罪活动和 β 表示犯罪行为. 如果 $\theta(C_i)>\alpha$, 那么给定的犯罪集团 C_i 就是活动的犯罪集团 A_i; 如果 $\Phi(C_i)>\beta$, 那么 C_i 就是一个严重犯罪集团. 如果一个犯罪集团既是严重的又是活动的组织, 我们就认为它是有组织犯罪集团, α 和 β 的具体取值的意义由实验结果决定.

3.2 有组织犯罪集团演化模型

这个模型需要确定原来的某个犯罪集团已经演变当前的某个犯罪集团. 一个犯罪集团的一个周期会出现五阶段: 产生、分裂、合并、出现和终止[10]. 为此, 引入一个匹配的函数:

$$match: Y\times 2^Y \rightarrow Y \tag{4}$$

这里的 g 表示一个犯罪组织集合, 2^g 表示 g 的幂集. 给定一个有组织犯罪集团 O_i^t 和有组织犯罪集团集合 g^{t+1}, 如果 $match(O_i^t,g)$ 得出集团 O_i^{t+1} 与 O_i^t 有最大的交集超过给定的阈值 l, $match(O_i^t,g)$ 形式定义如下:

$$O_j^{t+1}=O_k^{t+1}:O_k^{t+1}\in g \quad overlap(O_i^t,O_j^{t+1})\geq overlap(O_i^t,O_k^{t+1})\wedge overlap(O_i^t,O_j^{t+1})>l \tag{5}$$

这里两个有组织犯罪集团 $O,O'\in g$, $overlap(O,O')$ 定义如下:

$$overlap(O,O')=\min\left(\frac{|O\cap O'|}{|O|},\frac{|O\cap O'|}{|O'|}\right) \tag{6}$$

使用匹配函数, 应用下列规则对有组织犯罪集团演化进行跟踪:

1) 如果 $O_j^{t+1}=match(O_i^t,g^{t+1})$, 对每个 $O_k^{t+1}\neq O_i^t$, $O_j^{t+1}\neq match(O_k^t,g^{t+1})$, 那么 O_i^t 存在下一阶段 O_j^{t+1} 中;

2) 如果在这些分组与 O_i^t 之间有足够的重叠, 有 $(O_1^{t+1}$ 或 O_2^{t+1} ... 或 $O_n^{t+1})$ O_i^t 超过了预先给定的最小阈值, 那么 O_i^t 分成 $O_1^{t+1},O_2^{t+1},...,O_n^{t+1}$;

3) 如果 $O_j^{t+1}=match(O_i^t,g^{t+1})$ 且 $\exists O_k^t? O_i^t:O_j^{t+1}\neq match(O_k^t,g^{t+1})$, 那么由若干其他集团 O_j^{t+1} 合并成 O_i^t;

4)如果以上情况都没有出现, O_i^t 终止;

5)如果 $\forall O_i^t: O_j^{t+1}\neq match(O_i^t,g^{t+1})$, 那么 O_j^{t+1} 合并.

3.3 有组织犯罪集团挖掘算法

本文提出有组织犯罪集团挖掘算法主要包括:

1)计算犯罪网络矩阵共犯网络. 首先建立犯罪网络矩阵, 对得到的矩阵计算犯罪网络的犯罪次数得到子网络集.

2)筛选. 计算每个子网络的活动阈值和行为阈值, 利用操作算子得到活跃犯罪组织和严重犯罪组织.

3)挖掘犯罪核心组织. 对过滤后的网络结构, 评估出犯罪组织物质利益, 找出有组织犯罪组织的核心并预测组织的演变.

OCMM //计算犯罪核心, 预测有组织犯罪集团演变

输入:

1)犯罪事件数据集

2)Crime seriousness index(犯罪程度索引)

3)犯罪活动和犯罪行为的阈值: α,

输出: 有组织犯罪集团 $O_1^t,O_2^t,...,O_m^t$

步骤:

1:for each set of crime incidents in $[t_1,t_2]$\{

2: 计算犯罪网络矩阵 $N=MM^T$;// 用公式(1)操作算子计算

3: 计算犯罪组织 $C_1^t,C_2^t,...,C_n^t$ 的次数;

4: for each $C_i^t\in C^t$\{

5: 计算组织犯罪活动阈值 $q_i^{t_1,t_2}$ //用公式(3)操作算子计算

6: 计算组织犯罪行为阈值 F_i^t \}//用公式(2)操作算子计算

7: 对满足阈值的犯罪组织标识为有可能组织犯罪集团

8: for each possible organized crime O_i^t {

评估出犯罪组织物质利益}//用公式(4)操作算子计算

9: 对有组织犯罪集团O_1^t,O_2^t,…, O_m^t应用演变跟踪模型//用公式(5)和(6)操作算子计算}

}

4 实验结果分析

实验环境: i5 3.2GHz CPU, 2G 内存, 1TB 硬盘, Win8 操作系统, 所有程序使用 Java 语言在 JBuilder2008 开发环境下实现.

数据来源: 本实验采用加拿大不列颠哥伦比亚省 arrest-data 犯罪数据集[11], 该数据集包含所有报告犯罪信息(440 万条记录)及与犯罪有关的所有事件人(包括罪犯、受害者、目击者、起诉人和被指控人等), 总共有 39 个不同的主体组. 对于任何给定的犯罪事件, 每一个相关的主体有三个不同的状态字段, 说明主体在这一事件中的“角色”. 实验中我们限制受试者在这四个类别, 分别为控诉、可以控诉、推荐控诉或有嫌疑.

数据预处理: 从犯罪数据集中提取的共犯网络大约有 150,000 个结点和 600,000 条边, 其中结点度的平均大小为 4, 结点最大度为 525. 大约有 50%的结点度为 1, 这意味着这些罪犯在共犯网络中只有一次犯罪. 最大的连通网络部分大约占 18%结点链接组成, 这是非常大的网络结构部分. 实验中, 我们已经将数据集(按时间顺序排列)分为 5 个时间段, 每个时间段长为 12 个月. 为了考虑多个犯罪事件, 对 5 个时间段分别提取不同犯罪事件各自共犯网络, 对其识别罪犯组织、活跃的犯罪组织、严重的犯罪组织和有可能的有组织犯罪集团.

图1显示了使用规模大小为K的不同罪犯组织在5个时间段的数量. 正如所期望, 罪犯组织的数量随着群结构的增大而降低, 下面所有实验数据是对犯罪群的大小为 3 时进行讨论.

图 2 说明了不同犯罪行为阈值下严重罪犯组的数量. 大约 35%的罪犯组通过阈值 β= 0.5,这意味着有更大比例的罪犯组是小犯罪或轻度犯罪, 大约有 2%罪犯组通过阈值 β= 0.9, 说明只有很少的组织犯罪集团参与严重罪行.

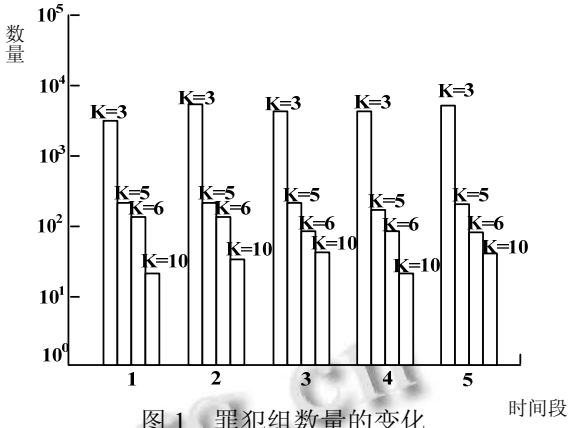

图 1 罪犯组数量的变化

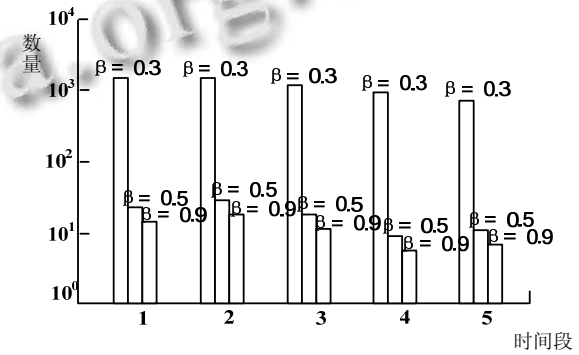

图 2 严重罪犯组数量的变化

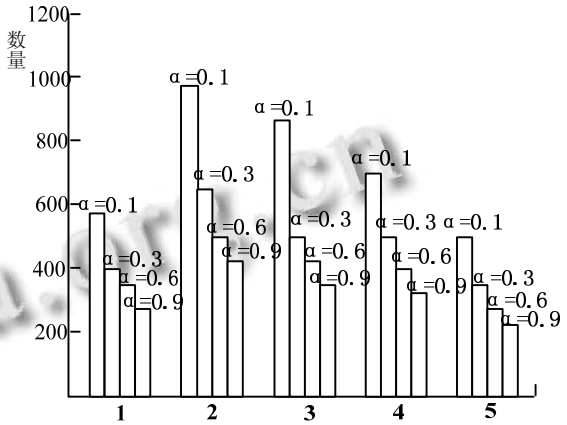

图 3 活跃罪犯组数量的变化

图 3 描述了在五个时间段活跃罪犯组数量的变化, 值得注意的是在图表中 α= 0.9 一直占罪犯组织的一半, 这意味着在较长时间一些犯罪组织保持的合作没有变化.

为了进一步讨论所有组织的数量, 将活动阈值 α 设为 0.3 犯罪阈值β设为 0.8, 在 五个时间段罪犯组、活跃组织和严重的组织的数量如图 4 所示.

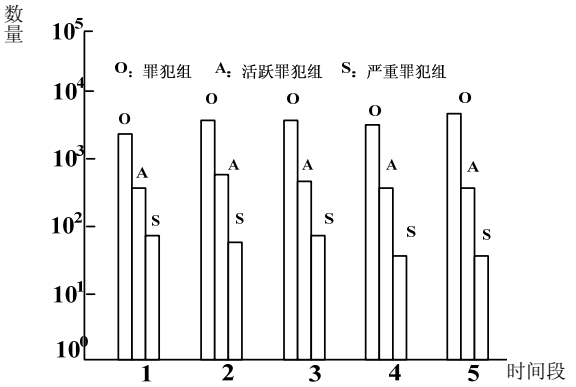

图 4　三种罪犯组数量变化比较

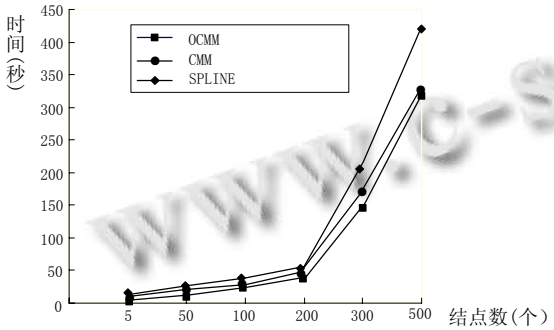

图 5　三种算法时间性能比较

从图 5 可以看出, 本文提出的 OCMM 算法和其它 CMM 算法、SPLINE 算法比较, OCMM 算法时间效率高.

最后对 35 组活跃和严重的犯罪组, 可以考虑他们可能是有组织犯罪集团. 有趣的是大多数犯罪组都有非常高的活动阈值, 这表明集团成员之间的关系非常紧密. 犯罪组织集的平均大小为 4.3, 这比严重和活跃犯罪组规模小, 这说明在增加集团犯罪数量时犯罪集团规模减少, 总体而言仅有 0.4%外围成员有可能成为有组织的犯罪集团成员, 内核成员并不急切与集团内核之外的犯罪成员进行合作.

我们的实验研究犯罪的最后一步是判断是否是有组织犯罪集团. 得出结论是, 有 7 组罪行是有限的性侵犯或骚扰犯罪有可能不与物质利益联系在一起; 有 6 组供认谋杀犯罪, 尽管他们可能非常严重有组织犯罪集团, 做出与物质利益相关的决定也相当困难; 其他 22 个组织承认有严重的犯罪, 包括绑架、药物与枪支走私、抢劫等等. 这些集团有很高概率成为有组织犯罪集团, 它们的成员也都离不开这个犯罪集团结构.

5　结语

控制有组织犯罪需要对犯罪网络进行调查, 然而犯罪团伙组成的复杂犯罪网络结构及隐蔽信息给公安机关和司法机构的执法带来很大困难. 为提高有组织犯罪集团挖掘算法的效率, 本文采用共犯网络矩阵得到犯罪集团的网络数据, 并在此基础上提出一种新的有组织犯罪集团挖掘算法 OCMM, 通过大量真实犯罪数据集实验验证了算法的有效性. 较大的有组织犯罪集团通常由若干个小的犯罪网络之间或较大犯罪网络之间交互而成, 虽然有组织犯罪集团挖掘算法能有效挖掘出集团核心成员, 但算法的运行时间仍然比较长, 算法的时间性能和检错性是未来进一步研究的工作.

参考文献

1 Brantingham PL, Ester M, Frank R, Glässer UM, Tayebi A. Co-offending network mining. In: Wiil UK, ed. Counterterrorism and Open Source Intelligence, Lecture Notes in Social Networks, Springer, 2011, 2: 211–239.

2 Morselli C. Inside criminal networks. Studies of Organized Crime. Springer 2009, 8: 103–120.

3 Reiss AJ. Co-offending and criminal careers. Crime and Justice. A Review of Research, 1988, 20(3): 25–34.

4 Fang MA. Analysis in the study of organized crime. Journal of Southwest University of Political Soience & Law, 2012, 14(2): 34–43.

5 Wen FL, Tang CJ, et al. Mining the core of crime network based on shortest path in social network analysis. Computer Science(S), 2006, 33(11): 266–268.

6 Liu QH, Tang CJ, et al. Mining the core member of terrorist crime group based on social network analysis. PAISI 2007, LNCS 4430. 2007. 311–313.

7 Newman MEJ. Fast algorithm for detecting community structure in networks. Phys. Rev. E, 2004, 69(3): 106–133.

8 Kim K, McKay R, Moon BR. Multiobjective evolutionary algorithms for dynamic social network clustering. Proc. the 12th Conf.Genetic and Evolutionary Computation. 2010. 1179–1186.

9 Inokuchi A, Washio T. Mining frequent graph sequence patterns induced by vertices. Proc. of the SIAM Int'1 Conf. on Data Mining. 2010. 466–477.

10 Girvan M, Newman MEJ. Community structure in social and biological networks. PNAS, 2002, 99(12): 7821–7826.

11 Tayebi MA, Glasser U. Organized crime structures in co-offending networks. Proc. Int'l Conf. on Social Computing and its Applications. Sydney, Australia. Dec. 2011. 1921–1954.