VCCM Mining: Mining Virtual Community Core Members Based on Gene Expression Programming*

Shaojie Qiao[1], Changjie Tang[1], Jing Peng[1],
Hongjian Fan[2], and Yong Xiang[1]

[1] School of Computer Science and Engineering, Sichuan University,
Chengdu 610065, China
{qiaoshaojie, tangchangjie}@cs.scu.edu.cn
[2] Department of Computer Science and Software Engineering,
the University of Melbourne, Australia
hfan@csse.unimelb.edu.au

Abstract. Intelligence operation against the terrorist network has been studied extensively with the aim to mine the clues and traces of terrorists. The contributions of this paper include: (1) introducing a new approach to classify terrorists based on Gene Expression Programming (GEP); (2) analyzing the characteristics of the terrorist organization, and proposing an algorithm called Create Virtual Community (CVC) based on tree-structure to create a virtual community; (3) proposing a formal definition of Virtual Community (VC) and the VCCM Mining algorithm to mine the core members of a virtual community. Experimental results demonstrate the effectiveness of VCCM Mining.

1 Introduction

Terrorist organization [1] is a complex adaptive system that emerged as an agent of change within the strategic system of nation states. It is an intricate network of individual small groups coupled by a common sense of purpose. So, how to distinguish the terrorists from the most likely suspects becomes a meaningful work.

Gene expression programming (GEP) [2, 3] is a new technique of evolutionary algorithm for data analysis. GEP combines the advantages of both GA and GP, while overcoming some of their individual limitations. If the attributes of classification samples are numeric, when applying GEP in the multi-dimension space classification, it will perform a global search in which genetic operators can select many attributes at a time. This paper makes the following contributions:

- Introduces a GEP-based classification algorithm to classify terrorists;
- Proposes a CVC algorithm to create a virtual community based on tree-structure;
- Presents an algorithm to mining the core members of a virtual community and experiments demonstrate that the searching cost can be reduced by this algorithm.

* This work was supported by National Science Foundation of China (60473071), Specialized Research Fund for Doctoral Program by the Ministry of Education (20020610007).

H. Chen et al. (Eds.): WISI 2006, LNCS 3917, pp. 133 – 138, 2006.

2 Related Work

Recently, there is a great deal of work about collecting terrorists' information and analyzing the system structure of the terrorist cell by computer [4, 5].

1. Searching the core of a terrorist group by computer. Jafar Adibi, a computer scientist at the University of Southern California, is developing ways to find hidden links between known terrorists and their as-yet-unknown confederates. He labels 20% of a terrorist group's members as "known" and challenges the program to find the rest.
2. Building a virtual al-Qaeda. Computer scientist Kathleen M. Carley heads a lab that tries to simulate terrorist organizations. The lab has built simulations of al-Qaeda by dumping newspaper articles into a computer database. A program then takes that information and looks for patterns and relationships between individuals.
3. Filtering data based on mathematical models. Mathematician Jonathan Farley of the Massachusetts Institute of Technology employs "order theory"—a branch of abstract mathematics that looks at the hierarchies within groups—to characterize the terrorist cells that intelligence agencies are trying to break up.

The differences between our work and these proposed methods lie in: comparing with related work one, this paper uses GEP to classify the terrorists, and treat classification as a means of data preprocessing; related work two builds a virtual al-Qaeda manually, but we uses computer to analyze and generate a virtual community.

3 Classification Algorithm Based on GEP

The preliminary concept is defined as follows.

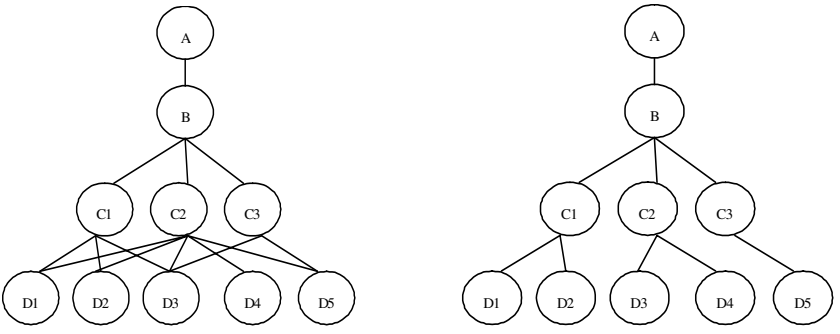
Definition 1 (GEP Classifier). A GEP classifier $C = (M, T, F, Op, v)$, where M is the set of GEP chromosomes [3]; T is the terminal set, such as {1, 2, a, b}; F is the function set, such as {+, -, *, /, sqrt}; Op is the set of genetic operators [6], i.e., selection, crossover, mutation and transposition [3]; v is the fitness value.

There are five steps in solving this problem by GEP (see Sect. 5.1 for detail).

1. to choose the fitness function;
2. to choose the set of terminals T and the set of functions F;
3. to choose the chromosomal architecture;
4. to choose the kind of linking function;
5. to choose the set of genetic operators and their rates.

4 VCCM Mining

In the communication perspective, al-Qaeda is much like a virtual community. The terrorist network can be described as Fig. 1 (a). The members of a virtual community are separated into two groups: *heads* and *members*. As shown in Fig. 1 (b), the nodes of the first three levels are heads, others are members. The nodes have the following features: (1) the nodes of the same level do not have edges linking with each other; (2) the weight of the edge is used to record the communication frequencies.



(a) A virtual community (b) A community tree

Fig. 1.

Definition 2 (Community Tree). A community tree is a 5-turple $T = (N, E, S, t, r)$, where N is a set of nodes; E is a set of undirected edges (u, v, c), where $u, v \in N$, u is the parent of v and c is the weight maps each node $n \in N$ to a set of values which is used to record the communication frequencies; S is a set of labels (C_h, C_i), C_h is the height of node n, and C_i is the sequence number of n in level C_h; t is a distinguished node called "taproot" of the tree; r is the root of the tree.

4.1 Creating a Virtual Community

There are four major steps in Create Virtual Community (CVC) algorithm.

1. Initialize the node set of the first three levels, i.e., {A, B, {C1,C2,…,Ci}};
2. Traverse each node (start from the fourth level) to find the nodes in the upper level which link with it and record all the communication frequencies (note, one node may have many edges joined by the nodes in the upper level), and let it be weight[i], where $i \in$ N; after that, compute each node's weights, let count = $\sum$ weight[i];
3. Sort the nodes of the same level by count in descending order using bubble-sort algorithm. In this way, one layer of nodes has been generated;
4. Repeat step 2-3 to create the following layer of nodes.

4.2 VCCM Mining

Definition 3 (Core Member of Community Tree). Let S[i] be the i-th set of leaf nodes which are siblings, and W be the weight sum of the nodes in S[i], $W=\sum_{j=1}^{k}$ weight[j], where k represents the node's number. S[i]′ is a subset of S[i], S[i]′⊆S[i], where S[i]′ = $\{n_{ij} \mid j = 1,2,\ldots,m$, where j represents the j-th node$\}$, $W'=\{\sum_{j \in M}$ weight[j]$\mid M \subseteq [1,k]\}$, $\eta=W'/W$, η is named *Factor*. If $\eta \geq \xi$ (ξ is a predefined threshold, where $\xi \in (0,1)$, and represents the significance of S[i]′), then n_{ij} is named Core Member of Community Tree and S[i]′ is the set of Core Members.

VCCM Mining has two phrases: *pruning* and *searching core members*.

The *Pruning* Phrase:

i) Compare the weights of node m_{ij}, where *i* represents the level number and *j* represents the *j*-th node (start from the fourth level), find the maximum, and save it;
ii) Delete other branches of node m_{ij};
iii) Use the same method to trim other nodes' branches until the last one.

The *Searching Core Members* Phrase: Let *T* be a community tree, where *r* represents the root and let *M* be the set of known core members (the node in *M* is *leaf*, and we use *M* to find other core members). First sort *r*'s children by weight in descending order, then create a stack S and push *r*'s children into it (note that, each sub-community which treats *r*'s child as its root at lest has one child in *M*), and sort these nodes by weight. Let W' be the weight sum of nodes in S, and *W* be the weight sum of nodes in *T*, calculate $\eta = W'/W$ and perform the following steps:

1) If $\eta \geq \xi$, go to step 3;
2) Otherwise, push the first node (*r*'s child) which is not in S into it, recalculate η; if η is still less than ξ, then push another node into it until $\eta \geq \xi$, go to step 3;
3) For each node *k* in S, if *k* is not a *leaf* node, repeat step1-2; if *k* is a *leaf* node , put it into a new stack S′ the nodes of which are core members;
4) Output the nodes in S′.

5 Experimental Evaluation

A model used to simulate the process for mining virtual community core members has been implemented in the developing platform of Microsoft Visual C++6.0. Experiments are conducted on a P4, 1.5 GHz PC with 256M RAM, running Microsoft Windows XP Professional. In order to validate the effectiveness of GEP-based classification algorithm, we use APS [7] to perform these experiments.

5.1 Experiment 1: Terrorist's Classification Problem

The data sets for classifying are synthetic data sets based on newspaper articles and other information about the terrorist organizations, and we named the terrorist database *Terrorist*. The training set contains 350 instances where the binary 1-bit encoding in which represents two possible output classes ("0" for false and "1" for true) and this database contains four testing sets: T1, T2, T3 and T4 which are discriminated by size shown in Table 1. Each instance is described by six attributes: *religion*, *origin*, *gender*, *is_educated*, *age* and *has_criminal_ records*.

The GEP function set includes {+, -, *, /, sqrt}, and the terminal set contains all the attribute names. In our experiments, the values of the parameters are: the head length is 8; the number of genes, chromosomes, and generations are 3, 100 and 100; crossover rate is 0.3; mutation rate is 0.044; and transposition rate is 0.1. Each experiment runs for ten times, and we use the average value to compare GEP approach with C4.5. Table 1 represents the test accuracy from these algorithms.

Table 1. Comparison of classification accuracy in terrorist's classification problem

Dataset	C4.5	GEP	Number of testing instances
T1	64.102	97.057	100
T2	64.062	96.628	200
T3	64.456	96.685	500
T4	62.678	96.571	1000

5.2 Experiment 2: Comparison of Efficiency Between GEP and GP

It is trivial to compare the CPU time between GEP with other traditional classification algorithms, i.e., C4.5, NaiveBayes and SMO [8], since GEP is definitely more time-consuming. But it is necessary to compare GEP with GP approach.

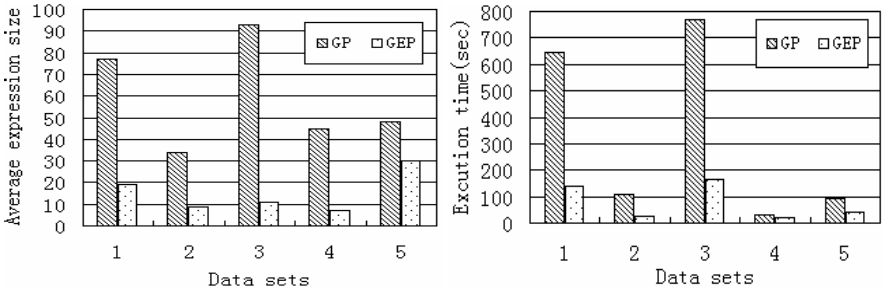


Fig. 2. Comparison of the average expression size and the execution time between GEP and GP

Figure 2 gives a comparison of the average expression size [2] generated by GEP and GP, and shows the comparison of the average execution time over ten different runs between GEP and GP on five benchmark data sets from UCI repository, i.e., *breast cancer*, *balance scale*, *waveform*, *zoo* and *iris*. It can be concluded that GEP tends to generate shorter expressions and costs less time compared with GP.

5.3 Experiment 3: Performance Evaluation of VCCM Mining

In Fig. 3 (a), X-axis represents the number of nodes generated by computer, and Y-axis is the *leaf* nodes (core members of a community tree), note that by using Normal Searching algorithm you have to search all *leaf* nodes.

Figure 3 (a) shows the cost for searching core members by VCCM Mining algorithm compared with Normal Searching algorithm. It is supposed that the time for searching each leaf node is equal, and experimental results demonstrate that the VCCM Mining algorithm can reduce the searching cost. Figure 3 (b) shows the changes of the searching cost with the number of nodes increasing. With the size of a community tree increasing, the VCCM Mining algorithm is still efficient.

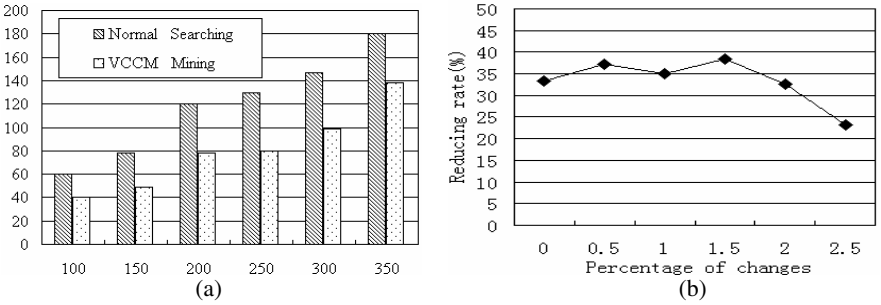


Fig. 3. Contrast of the cost of searching by VCCM Mining and Normal Searching algorithm

6 Conclusions and Future Work

In this paper, we analyze the characteristics of the terrorist organization and introduce a GEP-based classification approach. By using Community Tree, an algorithm is proposed to mining the VCCM. Experimental results show that VCCM Mining algorithm can reduce the cost for searching core members of a virtual community.

Our future work contains: extracting the important information about terrorists from Internet websites, because manual extracting is laborious and time-consuming; using the information to create the architecture of a terrorist group automatically; comparing the proposed algorithm with the traditional key network member identification methods such as network centralities and Carley's NETEST tool that combines multi-agent technology with hierarchical Bayesian inference models.

References

1. Larry K. Wentz and Lee W. Wagenhals: Effects Based Operations for Transnational Terrorist Organizations: Assessing Alternative Courses of Action to Mitigate Terrorist Threats. Proceedings of Command and Control Research and Technology Symposium, San Diego (2004)
2. Chi Zhou, Weimin Xiao, Peter C. Nelson, and Thomas M. Tirpak: Evolving Accurate and Compact Classification Rules with Gene Expression Programming. IEEE Transactions on Evolutionary Computation, Vol. 7, No. 6 (2003) 519–531
3. C. Ferreira: Gene Expression Programming: Mathematical Modeling by an Artificial Intelligence. Angra do Heroismo, Portugal (2002)
4. Matt Crenson: Math wizards offer help in fighting terrorism. http://www.azstarnet.com/dailystar/relatedarticles/42692.php (2004)
5. S. Qiao, C. Tang, Z. Yu, J. Wei, H. Li and L. Wu: Mining Virtual Community Structure Based on SVM. Computer Science, Vol. 32, No. 7 (2005) 208–212
6. J. Peng, C. Tang, J. Zhang and C. Yuan: Evolutionary Algorithm Based on Overlapped Gene Expression. ICNC 2005, Vol. 3612 of LNCS (2005) 194–204
7. C. Ferreira: Gene Expression Programming in Problem Solving. Soft Computing and Industry: Recent Applications, Springer-Verlag, Berlin Heidelberg (2002) 635–654
8. J. Platt: Sequential minimal optimization: A fast algorithm for training support vector machines. Advances in Kernel Methods-Support Vector learning, Cambridge, MA: MIT Press (1999) 185–208