Analysis of Telephone Call Detail Records Based on Fuzzy Decision Tree

Liping Ding[1], Jian Gu[2], Yongji Wang[1], and Jingzheng Wu[1]

[1] Institute of Software, Chinese Academy of Sciences, Beijing 100190, P.R. China
[2] Key Lab of Information Network Security of Ministry of Public Security
(The Third Research Institute of Ministry of Public Security),
Shanghai, 200031, P.R. China

Abstract. Digital evidences can be obtained from computers and various kinds of digital devices, such as telephones, mp3/mp4 players, printers, cameras, etc. Telephone Call Detail Records (CDRs) are one important source of digital evidences that can identify suspects and their partners. Law enforcement authorities may intercept and record specific conversations with a court order and CDRs can be obtained from telephone service providers. However, the CDRs of a suspect for a period of time are often fairly large in volume. To obtain useful information and make appropriate decisions automatically from such large amount of CDRs become more and more difficult. Current analysis tools are designed to present only numerical results rather than help us make useful decisions. In this paper, an algorithm based on fuzzy decision tree (FDT) for analyzing CDRs is proposed. We conducted experimental evaluation to verify the proposed algorithm and the result is very promising.

Keywords: Forensics, digital evidence, telephone call records, fuzzy decision tree.

1 Introduction

The global integration and interoperability of society's communication networks (i.e. the internet, public switched telephone networks, cellular networks etc.) means that any criminal with a laptop or a modern mobile phone may commit a crime, without any limitations on mobility [1]. There are more than 600 million cell phone users in China now. More and more frequently, investigators have to extract evidences from cell telephones for the case in hand. Telephone forensics is the science of recovering digital evidences from a telephone communication under forensically sound conditions using accepted methods. The information from CDRs includes content information and non-content information. Content information is the meaning of the conversation or message. Non-content information includes who communicated with whom, from where, when, for how long, and the type of communication (phone call, text message or page). Other information that is collected may include the name of the subscriber's service provider, service plan, and the type of communications device (traditional telephone, mobile telephone, PDA or pager) [2]. Once the law enforcement

X. Lai et al. (Eds.): E-Forensics 2010, LNICST 56, pp. 301–311, 2011.

agency obtains the telephone records, it may be important to employ forensic algorithm to discover correlations and patterns, such as identifying the key suspects and collaborators, obtaining insights into command and control techniques, etc. **Efficient and accurate** data mining algorithms are preferred in this case.

Software tools including I2's AN7 and our TRFS (Telephone Record Forensics System) are designed to filter and search data for forensic evidences. But these tools focus on presenting numerical analyzing results. The subsequent judgment, such as who is probably the criminal, who are probably the partners, and who has nothing to do with the event, will be made by the investigators based on their experiences. To address this issue, we propose a novel algorithm based on fuzzy decision trees to help the investigators make the final decision in this paper.

An investigator may analyze a suspect's telephone call records from two perspectives. One is global analysis in which we try to find all the relevant telephone numbers and their states that may be associated with a crimie incident. The other is local analysis in which we try to find a suspect's conversation content with someone and get important information. This paper focuses on the global analysis and tries to extract useful information (digital evidences) from non-content CDRs to help the investigator make decisions.

The rest of this paper is organized as follows. In Section 2, we introduce related work about telephone forensics, fuzzy decision trees, and our prototype of telephone forensics tool TRFS. We then present the algorithm based on fuzzy decision tree for CDR analysis in Section 3. In Section 4, we discuss our experimental evaluation and results. We conclude this paper and disucss future work in Section 5.

2 Related Work

2.1 Telephone Forensics

Mobile phones, especially those with advanced capabilities, are a relatively recent phenomenon, not usually covered in classical computer forensics. Wayne Jansen and Rick Ayers proposed guidelines on cell phone forensics in 2007 [3]. The guidelines focus on helping organizations evolve appropriate policies and procedures for dealing with cell phones, and preparing forensic specialists to contend with new circumstances involving cell phones. Most of the forensics tools that the guidelines proposed are designed to extract data from cell phones, and the function of data analysis is ignored. Keonwoo Kim, et al [4] provided a tool that copies file system of CDMA cellular phone and peeks data with an arbitrary address space from flash memory. But, their tool is not commonly applied to all cell phones since a different service code is needed to access to each cell phone and the logically accessible memory region is limited. I2's Analyst's Notebook 7(AN7, http://www.i2.co.uk) is a good tool that can visually analyze vast amounts of raw, multi-formatted data gathered from a wide variety of sources. However, AN7 is an aided tool for the investigator to find some patterns and relationships among suspects. Investigators have to reason themselves according to the

visual result derived from AN7. In this paper, we propose an algorithm based on fuzzy decision tree to help investigators infer and make their decisions more justified and scientific.

2.2 Fuzzy Decision Tree

The decision tree is a well known technique in pattern recognition for making classification decisions. Its main advantage lies in the fact that we can maintain a large number of classes while at the same time minimize the time for making the final decision by a series of small local decisions [5]. Although decision tree technologies have already been shown to be interpretable, efficient, problem independent and able to treat large scale applications, they are also recognized as highly unstable classifiers with respect to minor perturbations in the training data. In other words, this type of methods presents high variance. Fuzzy logic brings in an improvement in these aspects due to the elasticity of fuzzy set formalism. Fuzzy sets and fuzzy logic allow the modeling of language-related uncertainties, while providing a symbolic framework for knowledge comprehensibility [6]. There have been a lot of algorithms for fuzzy decision tree [7-11]. One of the popular and efficient algorithms is based on ID3, but it is not able to deal with numerical data. Several improved algorithms based on C4.5 and C5.0 have been proposed. All of them have undergone a number of alterations to deal with language and measure uncertainties [12-15]. The algorithms are not compared and discussed in details in this paper due to space limit. Our fuzzy decision tree algorithm for CDRs analysis introduce in the following is based on some of these algorithms .

A fuzzy decision tree takes the fuzzy information entropy as heuristic and selects the attribute which has the biggest information gain on a node to generate a child node. The nodes of the tree are regarded as the fuzzy subsets in the decision-making space. The whole tree is equal to a series of "IF…THEN…"rules. Every path from the root to a leaf can be a rule. The precondition of a rule is made up of the nodes in the same path, while the conclusion is from the leaves of the path. The detail algorithm is presented in Section 3.

2.3 Introduction of TRFS

TRFS is now only a prototype and have some basic functions as illustrated in Fig. 1 and Fig.2. It consists of six components: data preprocessing, interface, general analysis, data transform, special analysis, and others. CDR analysis is included in the special analysis as illustrated in Fig. 2. For example, utilizing CDR analysis, the investigators can carry out local analysis to find the telephone numbers that communicate with a suspect's telephone for less than N seconds, more than N seconds, or the earliest N telephone calls and the latest N telephone calls in a special day, etc.

TRFS has two important differences from AN7. AN7 does not only focus on telephone number analysis but also implement various kinds of analysis as financial, supply chain, projects, and so on. TRFS is a special system only for telephone forensics. Moreover, TRFS is based on Chinese telephone features and is suitable for Chinese telephone forensics. However, similar to AN7, TRFS can only give the

investigators numerical results and they have to make decisions based on their experiences. Therefore, we improve TRFS with fuzzy decision tree to support fuzzy decisions, e.g., who is probably the criminal, or who probably is the partner, etc.

Fig. 1. The main interface of TRFS

Fig. 2. The special analysis of TRFS

3 Proposed FDT Algorithm

A FDT algorithm is generally made up of three major components: a procedure to build a symbolic tree, a procedure to prune the tree, and an inference procedure to make decisions. Let us formally define FDT in the following. Suppose A_i (i=1,2,…,n) is the fuzzy attributes set of a training example data set D, $A_{i,j}$ (j=1,2,…,m) denotes the j^{th} fuzzy subset of A_i (m is different with different i.), and C_k (k=1,2,…,l) is the classified classifications.

Definition 1. (the fuzzy decision tree)
A directed tree is a fuzzy decision tree if

1) Every node in the tree is a subset of D;
2) For each non-leaf node N in the tree, all of its child nodes will form a subset group of D which is denoted as T. Then there is a variable k ($1 \leq k \leq l$), enables $T=C_k \cap N$;
3) Each leaf node is one or more values of classification decision.

Definition 2. (the rule of fuzzy decision tree)

A rule from the root to a leaf of a fuzzy decision tree is presented as:

If $A_1=v_1$ with the degree p_1 and $A_2=v_2$ with the degree p_2 …and $A_n=v_n$ with the degree p_n, then $C=C_k$ with the degree p_0 (1)

Definition 3. (the fuzzy entropy).

For a certain classification, suppose s_k is the number of examples from D in class C_k, the expected information can be calculated by

$$I(D) = -\sum_{k=1}^{l} p_k \log_2 p_k \tag{2}$$

where p_k is the probability of a sample belongs to C_k.

$$p_k = \frac{s_k}{|D|} \tag{3}$$

Defintion 4. (the membership function).

The membership values of the fuzzy sets are relevant to the edges of the tree. For the discrete attributes, classical membership function is usually adopted:

$$\mu_k = \begin{cases} 1, & if \quad d \in D_k \\ 0, & if \quad d \notin D_k \end{cases} \tag{4}$$

For continuous attributes, the trapezoidal function (5) and triangle function (6) are the popular membership functions.

$$\mu_k(x) = \begin{cases} 0, & x \le d_1 \\ \frac{x-d_1}{d_2-d_1}, & d_1 < x \le d_2 \\ 1, & d_2 < x \le d_3 \\ \frac{d_4-x}{d_4-d_3}, & d_3 < x \le d_4 \\ 0, & d_4 < x \end{cases} \tag{5}$$

$$\mu_k(x) = \begin{cases} 0, & x \le a \\ \frac{x-a}{b-a}, & a < x \le b \\ \frac{c-x}{c-b}, & b < x \le c \\ 0, & c < x \end{cases} \tag{6}$$

Also, the membership values of the fuzzy sets can be calculated through statistic methods by carrying out questionnaire among domain experts. Our algorithm is adopted (4), (5) and finally modified by invited computer forensics experts and investigators through statistic method.

After the generation of fuzzy decision tree, decisions can be made through inference. According to [16], the operator(+,×) among four kinds of operators(+,×), (V,×), (V,^), and (+,^) is the most accurately operator for fuzzy decision tree inference. Therefore, we use (+,×) to perform the inference.

3.1 Data Preprocessing

The raw data from telephone service providers is the telephone numbers and their detail records of outgoing calls or incoming calls of the suspect's telephone to be investigated. Several main attributes of the data we examine are *Tele_number*, *Call_kinds*, *Start_time*, *Location*, and *Duration*. The classes are *suspect*, *partner* and *none*. To fuzzify the data, we defined several sub attributes:

1) In *Call_kinds, call* and *called* present that the owner of the telephone *call*ed the suspect or was *called* by the suspect;
2) *early, in-day*, and *later* in *Start_time* denote the telephone conversation took place before, at or after the day that the crime is conducted;
3) *inside* and *outside* in *Location* present that the owner of the telephone was or was not in the same city (the region of a base station) with the suspect during their telephone conversation;
4) *long*, *mid* and *short* in *Duration* present the time spending on a telephone conversation.

All the definitions above are showed in Table 2.in Section 4.

3.2 Generation of Fuzzy Decision Tree

The key of generating a fuzzy decision tree is attribute expansion. The algorithm of the fuzzy decision tree generation in our system is as follows:

Input: Training example set E.

Output: Fuzzy decision tree.

Procedures:

For $e_g \in E$ $(g=1,2,\ldots p)$,
1) Calculate fuzzy classification entropy *I(E)*

$$P_k = \frac{\sum_{g=1}^{p} \mu_{gk}}{\sum_{k=1}^{l}\sum_{g=1}^{p} \mu_{gk}} \tag{7}$$

$$I(E) = -\sum_{k=1}^{l} p_k \log_2 p_k \tag{8}$$

where μ_{gk} is the membership of $e_g \in C_k$ ($g=1,2,\ldots p$, $k=1,2,\ldots l$).

2) Calculate the average fuzzy classification entropy of the i[th] attribute $Q_i(E)$

$$P_{ij}(C_k) = \frac{\sum_{e_g \in C_k} \mu_{gk}(A_{ij})}{\sum_{g=1}^{p} \mu_{gk}(A_{ij})} \tag{9}$$

$$I_{ij} = -\sum_{k=1}^{l} P_{ij}(C_k) \log_2 P_{ij}(C_k) \tag{10}$$

$$Q_i(E) = \sum_{j=1}^{m} \frac{\sum_{g=1}^{p} \mu_{gk}(A_{ij})}{\sum_{j=1}^{m} \sum_{g=1}^{p} \mu_{gk}(A_{ij})} I_{ij} \tag{11}$$

where $\mu_{gk}(A_{ij})$ is the membership of $e_g \in C_k$ under the attribute of $A_{i,j}$ (g=1,2,…p. k=1,2,…l).

3) Calculate the information gain.

$$G_i(E) = I(E) - Q_i(E) \tag{12}$$

4) Find i_0 which satisfies to

$$G_{i_0} = \max_{1 \le i \le n} G_i(E) \tag{13}$$

Then select A_{i0}as the test node.

5) For i=1,2,…,n, j=1,2,…m, repeat 2-4, until (1) the proportion of a data set of a class C_k is not less than a threshold θ_r , (2) there are no attribute for more classifications, then it is a leaf node and assigned by the class names and the probabilities.

3.3 Pruning Fuzzy Decision Tree

Pruning is to provide a good compromise between simplicity and predictive accuracy of the fuzzy decision tree by removing irrelevant parts in it. Pruning also enhances the interpretability of a tree. It is obvious that a simpler tree will be easier to interpret. Our pruning algorithm is based on [9], which is an important part of our method and will be discussed in detail in another paper in the future.

3.4 FDT Inference

As mentioned above, we adopted (+,×) to carry out the inference of the fuzzy decision tree. The algorithm is as follows:

Suppose the final fuzzy decision tree have v paths, every path has w_h nodes, the probabilities of the nodes is labeled f_{ht} ($h=1, 2, \ldots, v.\ t=w_1, w_2, \ldots, w_v.$). Every leaf node belong to C_k at the probability of $f_h^{C_k}$ ($k=1,2,\ldots l$)

Then

$$f_h^k = \prod_{t=1}^{w_h-1} f_{ht} f_h^{c_k} \quad (\text{h=1,2,}\ldots\text{v, k=1,2}\ldots\text{l}) \tag{14}$$

The total probability of classification is:

$$f^k = \sum_{h=1}^{v} f_h^k \tag{15}$$

And

$$\sum_{k=1}^{l} f^k = 1 \tag{16}$$

.

The reasoning formalization maybe:

If A_{h1} is Z_{h1} with the degree more than f_{h1} and A_{h2} is Z_{h2} with the degree more than f_{h2} and A_{hw_h} is Z_{hw_h} with the degree more than f_{hw_h} then $C = C_k$ with the degree f_h^k.

4 Experiment and Analysis

In a case of murder, we got the suspect's telephone number and collected 50 CDRs of some relevant telephone numbers during a period of time. Some of them are showed in Table 1. In the column of *Call_kinds*, 1 denotes the telephone *call*ed the suspect's telephone, while 0 denotes the telephone was *called* by the suspect's telephone. In the column of *Location*, every number presents the base station number which matches a certain geographic location. The time of the murder is about 2004/10/02 13:25:00. According to the algorithm in the above, the raw data is fuzzified and the membership is calculated by (4), (5). However, it is very complicated to determine which telephone owner is the main suspect, who is the partner and who has nothing to do with the event. For example, e23's telephone number is 114, which is the service provider of telephone number searching. So the owner of 114 may have nothing to do with the crime with a

high probability. In order to make the decision more accurate, we adopted a statistical method to imorve the calculated results. We invited 10 experienced investigators and 10 forensics experts to help us modify the membership values. The final result is illustrated in Table 2.

Using the data in Table 2 as the training example set and applying the method mentioned above, the entropies of the whole fuzzy set and the four fuzzy subsets are respectively:

$I(E)$=1.5685 $Q_1(E)$=1.8263 $Q_2(E)$=1.4830, $Q_3(E)$=1.5718, $Q_4(E)$=1.4146

Therefore the maximum information gain is *duration* and it is selected as the root node. The finally fuzzy decision tree is showed in Fig.3.

According to the inference method described in Section3, we can obtain the final probabilities of the three classes by operator (+,×) and get 21 rules from the fuzzy decision tree. For example, the path from the root to the left leaf node indicates 3 rules. One of them is:

If "Duration is short with the probability of more than 0.790" and "Start_ time is early with the probability of more than 0.443" then the owner of the telephone is suspect with the degree 0.473.

Following the rules derived from the FDT, investigators can determine the owner of an input telephone number is probably a suspect, or a partner, or has nothing to do with the case.

Table 1. Some of the original data

Telephone	*Call_kinds*	*Start_time*	*Location*	*Duration*
13061256***	0	2004/10/01 07:21:25	6	79
05323650***	0	2004/10/01 07:23:22	6	187
13605425***	1	2004/10/01 07:44:10	6	19
05324069***	0	2004/10/01 10:12:43	6	71
05324069***	0	2004/10/01 10:39:08	6	111
11*	0	2004/10/01 10:41:16	6	23
05322789***	0	2004/10/01 10:42:03	6	79
3650***	0	2004/10/01 11:59:02	6	69
13061256***	0	2004/10/01 13:44:36	6	120
13361227***	1	2004/10/01 14:03:51	6	35
13012515***	0	2004/10/01 17:36:00	6	50
13061229***	0	2004/10/01 17:37:23	6	20

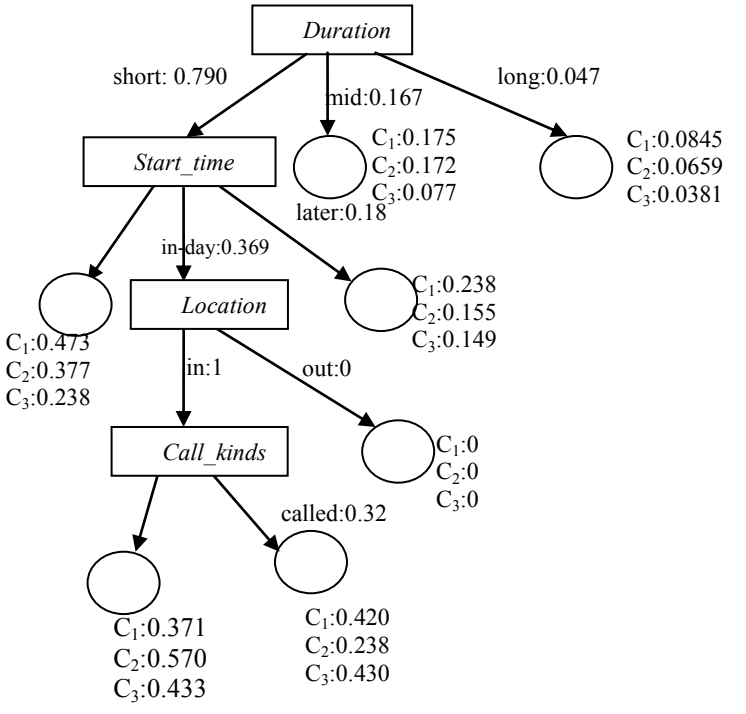

Fig. 3. The fuzzy decision tree

Table 2. Some of the original data

E	*Call_kinds*		*Start_time*			*Location*		*Duration*			*Classification*		
	called	call	early	in-day	later	In	out	short	mid	long	suspect	partner	non
e1	0.5	0.0	0.2	0.0	0.0	1	0	0.3	0.0	0.0	0.1	0.5	0.5
e2	0.5	0.0	0.0	0.3	0.3	1	0	0.6	0.0	0.0	0.2	0.6	0.3
e3	0.5	0.0	0.0	0.0	0.2	1	0	0.3	0.0	0.0	0.1	0.5	0.5
e4	0.0	0.3	0.0	0.0	0.2	1	0	0.3	0.0	0.0	0.3	0.4	0.5
e5	0.0	0.3	0.0	0.3	0.0	1	0	0.3	0.0	0.0	0.5	0.4	0.2
e6	0.5	0.3	0.6	0.6	0.0	1	0	0.6	0.0	0.0	0.5	0.5	0.1
e7	0.0	1.0	0.3	0.4	0.3	1	0	1.0	0.0	0.0	0.7	0.3	0.1
e8	0.5	0.0	0.0	0.3	0.0	1	0	0.3	0.0	0.0	0.1	0.7	0.5
e9	0	0.3	0.3	0.0	0.0	1	0	0.3	0.0	0.0	0.1	0.2	0.7
e10	0.0	0.3	0.3	0.0	0.0	1	0	0.0	1.0	0.0	0.3	0.3	0.4
e11	0.5	1.0	1.0	1.0	1.0	1	0	1.0	0.3	0.0	0.9	0.2	0.1
e12	0.6	0.0	0.6	0.0	0.0	1	0	0.6	0.0	0.0	0.3	0.5	0.4

5 Conclusions and Future Works

In this paper, we apply fuzzy decision tree to telephone forensics and enable investigators more justified reasoning. We discuss the related work of telephone forensics, FDT algorithms and our telephone record forensics system (TRFS). We then present our algorithm based on fuzzy decision tree. We further evaluate our algorithm with real experimental data. Currently, we are improving the algorithm by making FDT

generating, pruning and reasoning completely automatic, and looking into better methods to obtain appropriate membership values, and integrating the algorithm with our TRFS. In addition, the algorithm will be assessed and compared with other similar algorithms.

Acknowledgement. This research was supported by following funds: Accessing-Verification-Protection oriented secure operating system prototype under Grant NO.KGCX2-YW-125, the Opening Project of Key Lab of Information Network Security of Ministry of Public Security（The Third Research Institute of Ministry of Public Security）.

References

[1] McCarthy, P.: Forensic Analysis of Mobile Phones [Dissertation]. Mawson Lakes: School of Computer and Information Science, University of south Australia (2005)

[2] Swenson, C., Adams, C., Whitledge, A., Shenoi, S.: Advances in Digital Forensics III. In: Craiger, P., Shenoi, S. (eds.) IFIP International Federation for Information Processing, vol. (242), pp. 21–39. Springer, Boston (2007)

[3] Jansen, W., Ayers, R.: Guidelines on Cell Phone Forensics, http://csrc.nist.gov/publications/nistpubs/800-101/SP800-101.pdf

[4] Kim, K., Hong, D., Chung, K.: Forensics for Korean Cell Phone. In: Proceedings of e-Forensics 2008, Adelaide, Australia, January 21-23 (2008)

[5] Chang, R.L.P., Pavlidis, T.: Fuzzy decision tree algorithms. IEEE Trans. Syst. Man Cybern. SMC-7(1), 28–35 (1977)

[6] Zadeh, L.A.: Fuzzy logic and approximate reasoning. Synthese (30), 407–428 (1975)

[7] Quinlan, J.R.: Induction on decision trees. Machine Learning 1(1), 81–106 (1986)

[8] Doncescu, A., Martin, J.A., Atine, J.-C.: Image color segmentation using the fuzzy tree algorithm T-LAMDA. Fuzzy Sets and Systems (158), 230–238 (2007)

[9] Olaru, C., Wehenkel, L.: A complete fuzzy decision tree technique. Fuzzy Sets and Systems (138), 221–254 (2003)

[10] Umanol, M., Okamoto, H., Hatono, I., Tamura, H., Kawachi, F., Umedzu, S., Kinoshita, J.: Fuzzy decision trees by fuzzy ID3 algorithm and its application to diagnosis systems. In: IEEE World Congress on Computational Intelligence, Proceedings of the Third IEEE Conference on Fuzzy Systems, June 26-29, vol. (3), pp. 2113–2118 (1994)

[11] Kantardzic, M.: Data Mining Concepts, Models, Methods, and Algorithms. IEEE Press, Los Alamitos (2002)

[12] Ichihashi, H., Shirai, T., Nagasaka, K., Miyoshi, T.: Neuro-fuzzy ID3: a method of inducing fuzzy decision trees with linear programming for maximising entropy and an algebraic method for incremental learning. Fuzzy Sets and Systems (81), 157–167 (1996)

[13] Wehenkel, L.: On uncertainty measures used for decision tree induction. In: IPMU 1996 Info. Proc. and Manag. of Uncertainty in Knowledge-Based Systems, Granada, Spain (1996)

[14] Jeng, B., Jeng, Y., Liang, T.: FILM: a fuzzy inductive learning method for automated knowledge acquisition. Decision Support System (21), 61–73 (1997)

[15] Janikow, C.Z.: Fuzzy decision trees: issues and methods. IEEE Transactions on Systems, Man, and Cybernetics—Part B: Cybernetics 28(1), 1–14 (1998)

[16] Wang, X.Z., Yeung, D.S., Tsang, E.C.C.: A comparative study on heuristic algorithms for generating fuzzy decision trees. IEEE Transactions on Systems, Man and Cybernetics (31), 215–226 (2001)