ISBN 978-981-14-1684-2
Proceedings of 2019 the 9th International Workshop on Computer Science and Engineering
(WCSE 2019 SUMMER)
Hong Kong, *15-17 June*, *2019*, *pp. 580-587*
doi: 10.18178/wcse.2019.06.086

Research on Discovery and Classification Technology of Electric Power Marketing Field Terminals

Xianzhou Gao [1+], Ruxia Yang[1], Wei Chen[1], Congcong Shi[1]

[1] Global Energy Interconnection Research Institute, State Grid Key Laboratory of Information & Network Security, Nanjing, China

Abstract. In view of the characteristics of diversity, openness, complexity of Electric Power Marketing Field Terminals. There may be some security risks such as illegal terminal access. So the problem of discovery and classification of Electric Power Marketing Field Terminals is need to be solved, and then we can identify the types of illegal terminals in time and take corresponding measures. This paper aims at the diversity of Electric Power Marketing Field Terminals and their own differentiation characteristics, and proposes a technology without agent. Without installing client, it can automatically realize terminal discovery, and so solve the traditional problem of non-agent terminal discovery. At the same time, through K-means clustering algorithm terminal model identification, using unsupervised algorithm to extract and identify terminal type fingerprint information, it can achieve accurate classification of terminals, and provide timely alarm information and equipment data for network control and security protection.

Keywords: Unsupervised algorithm, K-means clustering algorithm, non-agent, accurate classification, terminal type fingerprint

1. Introduction

Electric Power Marketing Field Terminals such as marketing mobile operation terminal, charging POS machine, ATM automatic payment machine, video terminal, printer, fax machine and other common office peripheral terminal, have a variety of types, wide distribution, complex access mode, and have become an important entry point to implement network security attacks. Deep network attacks through Electric Power Marketing Field have become a difficult point of security protection. In order to effectively realize the safety management of Electric Power Marketing Field, firstly, the identification and discovery technology of Electric Power Marketing Field should be studied, and the effective discovery of terminal should be carried out. Then, the terminal of access network should be sorted out and analyzed comprehensively, and the access dynamics of terminal should be grasped in time. Secondly, the terminal classification technology should be studied to classify finely the discovered terminal and accurately identify the compliance terminal and the abnormal terminal. Further measures should be taken to provide the basis. However, for terminal recognition and discovery technology, it is difficult to deal with complex and diverse terminal types effectively through passive mechanism and traditional web service software. At the same time, there are a lot of supervised learning feature selection algorithms for terminal classification, and few feature selection algorithms for unsupervised learning. In this paper, an identification and discovery technology of Electric Power Marketing Field based on K-means clustering algorithm is proposed by combining active and passive methods, and unsupervised learning technology is deeply applied to extract and analyze fingerprints of terminal types, so as to achieve accurate classification of terminals.

2. Current Situation

+ Corresponding author. Tel.: + 025-83095549; fax: +025-83095588.
E-mail address: gaoxianzhou@163.com.

2.1. Research status of Technology

In recent years, network fingerprint identification technology has become a research hotspot in the field of network security. Scholars at home and abroad have put forward many theories and methods of network fingerprint identification. Reference [1] proposes a method of identifying Web server software by service identifier (Banner). Because the HTTP return packages of some terminal devices do not contain Banner information, this method has some limitations in identifying terminal devices. Literature [2] proposes a recognition method that does not rely on Banner information, i.e. identifying by the difference of some reason phrases returned and the difference of processing methods of ultra-long URLs. However, this method may increase the processing burden of terminal devices, cause denial of service, or be judged as an attack by devices such as firewalls, and trigger an alarm. In reference [3], a Web pattern recognition method similar to TCP/IP stack fingerprint recognition [4] is proposed. This method constructs malformed HTTP requests and identifies them according to the differences of different Web server software processing modes. However, the number of such differences is limited. The number of different types and models of terminal devices is much larger than the number of such differences. This method will lead to duplication and misjudgement. Document [5] sends 15 kinds of malformed HTTP requests to Web servers, and uses the returned status codes as input, constructs naive Bayesian classifier to classify the mainstream Web server software, but fails to recognize other terminal devices such as wireless routers, IP cameras, intelligent switches and so on.

In order to categorize terminals, machine learning is needed. Machine learning algorithms can be divided into supervised learning and unsupervised learning. Supervised learning refers to training samples with labels, while unsupervised learning does not have labels in the training process. In the real world, most samples are unlabeled, so unsupervised learning is more widely used than supervised learning. The commonly used unsupervised learning algorithms include PCA [6], isometric mapping [7], local linear embedding [8], Laplace feature mapping [9], Hesse local linear embedding [10] and local tangent space arrangement [11]. The problem of feature selection in unsupervised learning is to select a feature subset which can best cover the natural classification of data according to certain criteria. Current methods include feature selection method based on genetic algorithm [12], feature selection method based on pattern similarity judgment [13] and feature selection method based on information gain [14]. These methods do not consider the correlation between features and the impact of features on classification. Document [15] proposes an unsupervised feature selection method. The basic idea is: firstly, the samples are classified by competitive learning algorithm to determine the number of classifications; secondly, the original feature set is divided into several feature subsets, and the judgment function J=trace($(\sum C + \sum S)^{-1} \sum S$) is calculated in each feature subset, where $\sum C$, $\sum S$ denotes the average dispersion within the class, respectively. At last, the correlation coefficient between candidate features and selected features is calculated. If the correlation coefficient is greater than 0.75, the candidate features are discarded.

2.2. Electric power marketing field terminals demand and problems

According to statistics, the number of electric power marketing field terminals is the largest, accounting for 84.87% of the total terminals. There are nearly 10 types of marketing sites deployed in the marketing field, including charging POS, ATM automatic payment machines, call machines, office computers, scanners, video surveillance terminals and so on. There are various terminal access modes, communication protocols, business applications, operating systems, etc. All kinds of terminals are mainly controlled by simple IP+MAC address binding, and power self-service payment terminals. The power wireless POS terminal and the power cable POS terminal are specialized marketing terminals, which use the communication protocol of the integrated payment management platform to interact with the main station.

Due to the variety, diversity and complex business scenarios of marketing field terminals, the current security access measures are difficult to cover, and some non-traditional terminals such as cameras, call machines and printers cannot install clients or transform to achieve access control, and cannot achieve unified security control. First, the terminal site environment is uncontrollable and vulnerable to violent destruction for illegal network access. Compared with other terminals, the marketing field terminal mainly provides services to the public, deployed in an open site environment, facing a complex physical and personnel environment. At the same time, due to the large number of marketing field terminals and the lack of centralized deployment, it has become the "preferred point" for illegal personnel to carry out network

attacks. Second, terminal security protection measures are not perfect, illegal personnel can access the network by imitating terminals. At present, most of the marketing field terminals adopt IP + MAC address binding to access control. Illegal personnel easily obtain IP, MAC address and other information through "network sniffing" and "ARP spoofing", and then imitate IP, MAC address for illegal access. Thirdly, the software and hardware platforms of marketing field terminals are diverse, and some terminals cannot install clients or carry out modifications to achieve access control. Because of the need of marketing business, there are many types of terminals, such as toll POS, ATM automatic payment machine, call machine, office computer, scanner, video surveillance terminal, etc. These terminals have great differences in access mode, communication protocol, business application, operating system and so on. It is impossible to install Terminal agent like traditional desktop office terminal or to improve it through hardware transformation. Access control. Therefore, in view of the uncontrollability of the marketing field terminal environment, there is an urgent need to study the access control technology of the marketing field terminal. Firstly, the access terminals should be found in time. Secondly, the types of terminals should be sorted out and categorized so as to lay a foundation for the safety management and control of the marketing field terminal.

3. Design of Electric Power Marketing Field Terminals discovery in marketing Field

Scanning detection is carried out through TCP to identify and determine the operating system type and version of the terminal according to the difference of characteristic information between TCP/IP stacks of different devices [16-19]. According to the four-layer protocol stack of TCP/IP, including TCP, IP, UDP, ICMP, ARP and data link layer protocol, the process implements the application layer protocol. According to all the corresponding settings in the fingerprint database, the protocol stack parses the protocol data of each layer of TCP/IP, and decides whether to discard, continue to operate, reconnect and respond after judging. The protocol is encapsulated according to the fingerprint database, and the domain setting of each protocol header and the content of the response information are determined. According to the differences in the characteristics of TCP/IP stacks of different hosts. The terminal feature information can be identified and determined. In order to capture the original data packet in high-speed network environment, reduce the rate of packet loss and improve efficiency, protocol filtering is carried out in the process of packet capture.

As shown in Figure 1, the basic flow of the scheme is divided into three layers.

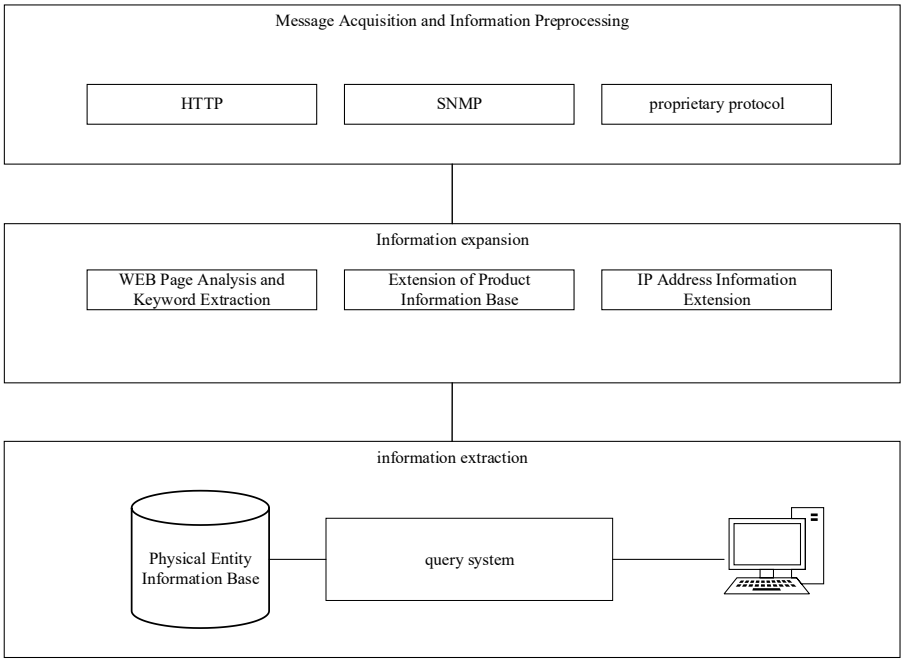

Fig. 1: Flow chart of terminal identification processing.

The first layer is message inquiry and protocol pre-processing layer. The main work is to detect and analyze HTTP, SNMP, communication protocol of integrated payment management platform for a specific IP. The second layer is the information expansion part, which mainly enriches the information acquired in

the next step, for example, key eigenvalue extraction for return data of protocols such as HTTP/SNMP/Integrated Payment Management Platform Communication Protocol; for some devices that can acquire specific models, detailed parameters about device hardware can be obtained by model matching; for IP/MAC address information, it can be obtained. The topological location of the device is taken to further expand the information of the device. In the third layer, the information extracted from the return information of the device and the information obtained from the extended analysis will be extracted and stored in the corresponding database.

4. Research on Categorization Technology of Electric Power Marketing Field Terminals

4.1. Classification and recognition process design

On the basis of automatic terminal discovery and collection of terminal equipment information and network protocol features, K-means network space terminal equipment identification model based on cosine measure is adopted. The model uses unsupervised learning clustering method to eliminate the calculation of prior probability of training set, and identifies different types of terminal equipment from different manufacturers and different versions of terminal equipment, thus realizing terminal return in marketing field. Category provides the basis for terminal access control in marketing field. The classification technology based on pattern recognition consists of three stages: data preprocessing, classification analysis and matching decision. In the data pre-processing stage, firstly, the extracted equipment information and network information are fused and feature extraction is carried out to obtain feature set; furthermore, in the classification analysis stage, training samples are analyzed and corresponding classifiers are constructed, and on this basis, the feature set and classification of equipment are carried out; in the matching decision stage, the final classification is obtained by pattern matching. Class results. Figure 2 shows the specific process:

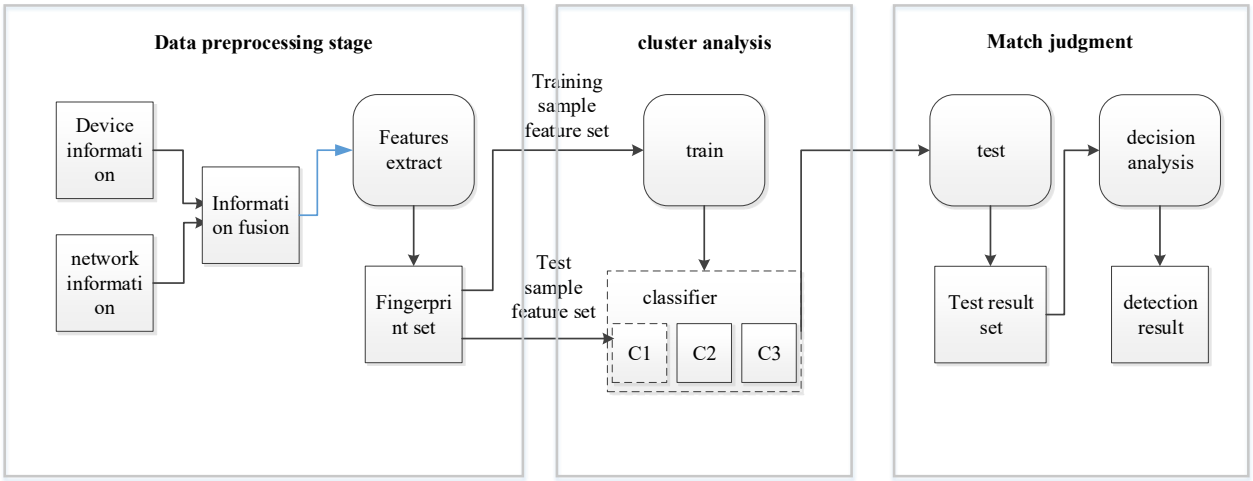

Fig. 2: Terminal Classification Technology Based on Device Type Fingerprint

4.2. K-means Clustering Model

By dividing the sample set into several clusters according to the similarity criterion, the clustering effect of similarity within clusters and difference among clusters is finally achieved 13. In the process of clustering, K is first determined manually, and K samples are randomly selected as the initial clustering centers in the sample set; the similarity between each sample set and all clustering centers is calculated and divided into the most similar clusters; and then the average value of each cluster is recalculated as the new clustering centers. The whole process is repeated until the clustering criterion function converges.

The algorithm steps are as follows:

1) Let the given sample X = $\{x_1, x_2, x_3, \ldots, x_n\}$, where x_i is the d-dimensional eigenvector of the first sample. Given the number of clusters K, K initial clustering centers are randomly selected in the sample set X, and are recorded as $c_1, c_2, \ldots, c_K$.

2) For n samples of a given sample set, the similarity degree between them and the cluster centers is calculated according to the similarity measure function, and the similarity degree is divided into K clustersC_1, C_2, … , C_j.

3) Calculate the average value of each cluster as a new clustering center.

4) Computing Clustering Criterion Function,

$$J = \sum_{j=1}^{K} \sum_{x_i \in C_j} d(x_i, c_j), \tag{1}$$

In the formula, $c_j$denotes the cluster center of clusterC_j, and $d(x_i, c_j)$ denotes the similarity measure function.

5) If the clustering criterion function converges, the algorithm will be terminated; otherwise, steps 2 to 4 will be repeated until the clustering criterion function converges.

6) When the clustering criterion function converges, the J value is the smallest and the clustering effect is the best.

4.3. Unsupervised feature selection algorithm based on K-means

For each sample set F_i, we use K-Means clustering algorithm to cluster the samples and determine the corresponding clustering number k_i. DB Index criterion is used to judge the clustering validity. Given a sample set X, clustering is carried out without any information of sample distribution. The optimal number of clusters will not exceed $k_{max} = \sqrt{n}$ [20] by iteration. Therefore, the iterative algorithm can be carried out between $k_{min} = 2$ and $\sqrt{n}$, and we can set a k_{max} value far less than $\sqrt{n}$ according to the specific application. The process of determining the clustering number k_i is as follows:

1) Initialization, C = 2, DB* = infinity, k_i= 1. Among them, C is the iteration variable of the number of classes, k_i is the best number of classes, and DB* is the smallest value of DB.

2) Clustering the samples by K-Means clustering algorithm. We establish a judgment function as shown in formula (2). When $d_j(i) \leqslant \alpha$(α is a set threshold), the clustering ends, and DB_c=$DB_c(i)$.

$$d_j(i) = \frac{|DB_c(i+1) - DB_c(i)|}{DB_c(i)} \tag{2}$$

Among them, $DB_c(i)$denotes the value of the second clustering DB whose clustering number is C.

3) If DB* < DB_c, then DB* =$DB_c(i)$, k_i= C.

4) C = C + 1, if C < k_{max}, the operation will continue, otherwise the clustering will end. k_i is the best classification number corresponding to the first feature subset.

Two subsets of features F_i, F_j (i = 1... t, j=1... t, i, j and t is the number of feature subsets) corresponding features are not exactly the same, so for different feature subsets F_i, the values of DB_{k_i} and DB_{k_j} obtained by F_j are not directly comparable, so it is necessary to standardize the judgment rules. Assuming the corresponding classification result C_i of F_i, the judgment function is

$$\text{Crit}(F_i, C_i) = DB_{k_i} \tag{3}$$

Then Crit(F_j, C_i)=DB is obtained by using the classification result C_i in the F_i feature subset. Then a standard judgment function is defined as shown in Formula (4). The selection of feature subset is to select the F_i with the smallest causation (4).

$$\text{Norma lizedcrit}(F_i) = \frac{1}{t} \sum_{p=1}^{t} \text{Crit}(F_i, C_i) \tag{4}$$

In the literature [21, 22], it is proposed that it is better to select the best feature subset than to select the best feature subset to form the feature subset. Therefore, we use sequential deletion method to search the feature subset in the algorithm. Let F be the original feature set and the feature dimension m, let t = m, count = 1, normal = 0, where t records the number of feature subsets, count records the number of execution times of the algorithm, and normal stores the value of norma lizedcrit of the best feature subset selected previously. The basic steps of the algorithm are as follows:

1) Delete one feature x_i from it in turn, and get t feature subsetsF_i, i=1... t. The optimal classification number k_i of these feature subsets is obtained by using the above methods.

2) The judgment rule of selecting feature subset is used to select the minimal F_i of causation (4). t=t-1, F =F_i.

3) If |norma lizedcrit(F_i) - normal|> β （β is pre-set threshold）and count≤m, so normal =norma lizedcrit(F_i), count = count+1.

4) The feature correlation of the selected feature subset F_i is analyzed. If the correlation coefficient of the two features is greater than γ（γ is the threshold), one of the features is deleted.

5. Categorization Results of Electric Power Marketing Field Terminals

Through the application of marketing field terminal, devices in the network are found and classified. The unsupervised algorithm based on K-means clustering can quickly find the whole network devices, and realize the precise classification of terminals with operating system, cameras and printers.

Fig. 3: The classification results of electric power marketing field terminals.

IP	MAC	Finger print	Equipment	Locked	User	Type
192.168.44.119	F0:7B:CB:A6:7E:4D		PEDONE	--	--	PC
192.168.44.165	E4:CE:8F:23:0B:AA		192.168.44.165			PC
192.168.44.103	E4:25:E7:81:CA:A3		192.168.44.103			PC
192.168.44.189	D4:3D:7E:7D:17:D1		192.168.44.189			PC
192.168.44.7	CC:B2:55:9D:16:82	--	192.168.44.7			PC
192.168.44.95	CC:AF:78:2A:19:96		192.168.44.95			PC
192.168.44.100	C0:63:94:D1:0C:C7	--	192.168.44.100	--	--	PC
192.168.46.168	B8:88:E3:EC:3C:02		PC201307060920	--	--	PC
192.168.46.142	B0:51:8E:02:46:E7		192.168.46.142	--	--	PC
192.168.44.115	A4:17:31:F6:EF:98		192.168.44.10	--	--	PC
192.168.44.115	A4:17:31:F6:EF:98		YU-PC	--	--	PC

Fig. 4: The type fingerprints of electric power marketing field terminals.

6. Concluding Remarks

Electric Power Marketing Field Terminal environment is complex, various types, a large number of, the private construction of terminals, illegal access and other issues have always been the difficulty of security management. This paper studies the technology of terminal discovery and classification in marketing field. By deploying related systems in bypass, relying on agent mode to report information and active detection and discovery mechanism, it achieves the discovery ability of active and passive combination. On the basis of designing unsupervised K-means clustering algorithm, it can effectively deal with traditional PC terminals, cameras, printers and other terminals that cannot install agents. The problem of discovery and classification of terminal equipment has good application value in marketing field terminal and Internet of Things terminal, which provides a basis for further development of security protection of power marketing field terminal access network.

7. Acknowledgements

This paper was financially supported by the science and technology project of State Grid Corporation of China: "Research on Key Technologies of Marketing Site Terminal Security Access" (Grand No. SGGR0000XTJS1800079).

8. References

[1] Shah S. An introduction to HTTP fingerprinting [EB/OL]. (2004-05-19) [2015-12-31]. http:// net-square.com/httprint_paper.html

[2] Lee D, Rowe J. Ko C, et al. Detecting and defending against Web-server fingerprinting [C] // CSAC 2002: 2002 Computer Security Applications Conference. United States: IEEE Computer Society,2002: 321-330.

[3] Kexin Yang, Jiubing Ju. Service Mapping Using Web Fingerprint[J]. Computer Engineering and Application, 2004,40(4) : 7-9.

[4] Fyodor. Remote OS detection via TCP/IP stack fingerprinting[J].Phrack Magazine, 1998, 17(3) : 1-10.

[5] Shaohua Wu, Dan sun, Yong Hu. Web Server Recognition Based on Bayesian Theory[J].Computer Engineering,2015,41(7) : 190-193,198.

[6] Jose CA. Fast On-line alogrithm for PCA and its convergence characteristic. IEEE Trans. on Neural Network, 2000, 4(2): 299–307.

[7] Tenenbaum JB, de Silva V, Langford JC. A global geometric framework for nonlinear dimensionality reduction. Science, 2000, 290(5500): 2319–2323.

[8] Roweis ST, Saul LK. Nonlinear dimensionality reduction by locally linear embedding. Science, 2000, 290(5500): 2323– 2326.

[9] Belkin M, Niyogi P. Laplacian eigenmaps for dimensionality reduction and data representation. Neural Vomputation, 2002, 15: 1373–1396.

[10] Donoho DL, Grimes C. Hessian eigenmaps: Locally linear embedding techniques for high dimensional data. PNAS, 2003, 100(10): 5591–5596.

[11] Zhang Z, Zha HY. Principal manifolds and nonlinear dimensionality reduction via tangent space alignment. SIAM Journal on Scientific Computing, 2004, 26(1): 313– 338.

[12] M M orita, R Sabourin , et al . Unsupervised Feature Selection Using M ultiObjective Genetic A lgorithm s forHandw rittenW ord Recognition [C]. Edinburgh , Scotland: International Conference on DocumentAnalysis and Recognition (ICDAR’ 03), 2003 . 666-671.

[13] Jayanta Basak, RajatK De, SankarK Pa.l Unsupervised Feature Sel ectionU sing a Neurof uzzy Approach[J]. Pattern Recognition Lett ers , 1998, 19(11): 9971006.

[14] M Dash , H Liu, JYao . D imensionality Reduction ofUnsuperv ised Dat a[C]. Newport Beach: Proc . 9th IEEE Int' lCon. f Toolsw ith A rtifical Intelligence , 1997. 532539.

[15] N icolasV, L M, JG Postaire . Unsupervised Co lorTexture Feature Extraction and Selection forSoccerImage Segmentation[C]. Vancouver , Canada: IEEE International Conference on Image Processing (ICIP ’ 2000), 2000. 800-803

[16] Zhengming Ma. The Principle and Application of TCP/IP. Beijing: Metallurgical Industry Press,2006.

[17] Ying Liu, Zhi Xue, Yijun Wang. Recognition of Remote Operating System Based on TCP Protocol Options. Information security and communication secrecy,2007;(11) :71-72.

[18] Chao Sha, Yunfang Chen. An Operating System Recognition Technology Based on TCP/IP Protocol Stack.Computer Technology and Development, 2006;(16) :125-128.

[19] Douglas E. Comer, Yaoyi Lin. Network Interconnection with TCP/IP. Beijing: Electronic Industry Publishing House,2001.

[20] Jian Yu, Qiansheng Cheng. Search Range of Optimum Cluster Number in Fuzzy Clustering Method [J]. Science in China, 2002, 32(2): 274280.

[21] Kira k, LA Rendell A Practical Approach to Feature Selection[C]. The 9th International Conference on Machine Learning, Morgan Kaufmann , 1992 . 249256.

[22] Elashoff J D, et al . On the Choice of Variables in Classification Problem swith Dichotomous Variables[C]. Biometrika , 1967. 668770.