Research on characteristic parameters mining and clustering of unknown protocols bitstreams

Yang Wu*, Tao Wang, Jin-dong Li

Dept. of Information Engineering, Shijiazhuang Mechanical Engineering College, Shijiazhuang, 050003, P.R. China

**Corresponding author's e-mail: baiyanwy@163.com*

Received 25 November 2014, www.cmnt.lv

Abstract

Characteristic parameters mining of unknown protocol bitstreams and parameters optimizing of clustering algorithm are the foundations of unknown protocol bitstreams analyzing. The parameters such as the bit frequency, runs and bit frequency within a block are defined according to the frequency of zero and one, frequency of sequential zero and one, bit frequency within a block. As the parameter of bit frequency within a block is sensitive to the block length, an optimal block length selection algorithm is proposed based on the principle of variance. In order to select effective initial clustering centers for division clustering algorithms such as the k-means algorithm, an initial clustering centers selection algorithm is proposed based on the peak value of sample density for each dimension. In order to select the optimal clustering number, a function of clustering quality evaluation is given by the sample density in cluster and cluster density. Taking the bitstreams of HTTP, DNS, ICMP, TELNET and UDP datasets as the unknown protocols bitstreams, the experimental results not only verified the effectiveness of the proposed algorithms but also point out the necessity of mining more effective parameters.

Keywords: Unknown protocol, bitstreams, clustering, characteristic parameter, bit frequency within a block

1 Introduction

Generally Speaking, the main task of unknown protocol identification is to find out the format information of the protocol from its bitstreams based on frequent sequences mining and the established association rules. Unknown protocol identification can provide supports for further unknown protocol analysis and utilization. Dividing the unknown protocol bitstreams with similar characteristics into corresponding clusters is the foundation of frequent sequences mining and unknown protocol identification. As the known protocol data is the main research object of protocol identification currently, the protocols of network data are distinguished mainly based on pattern matching [1], machine learning [2] and some other known protocol identification methods. The network data capture and analysis tools such as the Snifter and Ethereal are all based on above methods. The main challenges of unknown protocol identification are to identify the protocol fields of unknown protocol and user data accurately in the absence of any prior knowledge of the unknown protocol. However, most of the existing protocol identification technologies are based on the characteristics of known protocols; they cannot be effectively used to analyze the unknown protocol data.

2 Related works

It is easy to know that the key processes of unknown protocol bitstreams clustering mainly include characteristic parameters mining of bitstreams, initial clustering centers selection, optimal clustering number selection, clustering results evaluation and other key issues.

2.1 CHARACTERISTIC PARAMETERS MINING OF BITSTREAMS

The traditional characteristic parameter mining contents of the bitstreams include protocol type, ports, bitstream length, bitstream direction, characteristic fields and some other characteristic parameters. In 1999, the MIT Lincoln Labs provided the 41-dimensional real network traffic data for KDD competition, which is the acknowledged DARPA data for intrusion technologies testing [3]. But to unknown protocol characteristic parameter mining, there are only a few relevant researches in encrypted traffic identification. Charles [4] proposed a method to identify the application protocol of encrypted traffic according to the bytes number of data packets, durations, interactive processes and flow directions. Based on the interactive processes of SSL/TLS traffic, Sun [5] proposed a hybrid multi-level encrypted SSL/TLS traffic classification method, which identifies the specific application protocol of the encrypted traffic by statistical analysis. From the perspective of protocol independent, literature [6] provided a protocol independent online identification scheme for encrypted traffic by extracting the different statistical information of the encrypted and non-encrypted bitstreams. Literature [7] also proposed an encrypted bitstream identification scheme based on the statistical distributions of the zero and one in random and un-random bitstreams.

2.2 CLUSTERING ALGORITHMS AND THEIR PARAMETERS SELECTION

(1) Clustering algorithms selection

Clustering is one of the most important data analysis methods; it divides the samples with similar attributes into corresponding clusters according to a certain similarity measure rule. However, all of the clustering algorithms cannot be widely used to reveal the structures of multidimensional data [8]. The traditional clustering methods mainly include division clustering, hierarchical clustering, grid-based clustering, density-based clustering and model-based clustering. Most of the clustering algorithms are sensitive to their parameters; different parameters may bring completely different clustering results. As division clustering methods have lower implementation complexity, they are widely used in large-scale data clustering; many researchers naturally pay their attentions to the research of parameters selection for division clustering. There are many typical division clustering algorithms such as the *k*-means, PAM (Partitioning Around Medoid), *k*-modes and EM (Expectation Maximization) algorithm [9]. The pivotal problems of the division clustering algorithms mainly include initial clustering centers selection, optimal clustering number selection and clustering results evaluation.

(2) Initial clustering centers selection

The initial clustering centers selection methods of division clustering algorithm mainly include the RS (Random Selection) method, MMD (Maximum and Minimum Distance) algorithm and other improved algorithms.

(a) RS method: If the number of clusters is *k*, the RS method randomly chooses *k* samples as the initial clustering centers. Although the process of the RS method is very simple, its clustering results are usually inconsistent. Different initial clustering centers could inevitably result in different clustering results.

(b) MMD algorithm: The basic idea of the MMD algorithm is to select the samples with maximal distance as the initial clustering centers.

To avoid clustering algorithm converging to a local minimum, Likas [10] proposed a global *k*-means algorithm, in which the initial clustering centers are more and more close to the real clustering centers during the iterative processes. In order to increase the likelihood of obtaining the globally optimal solution, literature [11] provided an initial clustering centers selection algorithm based on selecting the dispersed samples as the initial clustering centers. Based on the MMD algorithm, literature [12] proposed a scheme to select the high-density points farthest from the initial clustering centers as the new centers. Literature [13] proposed a fuzzy clustering algorithm based on large density region to avoid the clustering algorithm converging to a local minimum, but the algorithm needs to calculate the density values of all samples, it is not suitable for large-scale data clustering. Literature [14] proposed a method using recursive calls to find the initial clustering centers with farthest distance for the *k*-means algorithm.

(3) Optimal clustering number selection

The optimal clustering number has important significance for getting high accuracy clustering results. Many classical indices are proposed such as the CH (Calinski-Harabasz) index [15], DB (Davies-Bouldin) index [16], KL (Krzanowski-Lai) index [17], Wint (Weighted inter-intra) index [18], IGP (In-Group Proportion) [19] and so on. But all of these indices are often unable to obtain the correct clustering number when the clustering structures are difficult to determine. Literature [3] proposed a clustering results evaluation method called COPS (Clusters Optimization on Preprocessing Stage) based on hierarchical division, which effectively improves the accuracy of clustering number selection. Furth more, literature [9] proposed the BWP (Between-Within Proportion) index for the *k*-means algorithm. All of the above indices are based on Euclidean distances of the samples or clusters, with the increase of sample dimension, the distance approaching phenomenon will be more obvious and the above methods will become invalid.

3 Unknown protocol bitstreams clustering scheme

3.1 CHARACTERISTIC PARAMETERS MINING FOR UNKNOWN PROTOCOL BITSTREAMS

(1) Bit frequency statistics parameter mining

Firstly, the bitstreams are $DB=(X_1,X_2,...,X_N)$, where $X_i=\left(x_1^i,x_2^i,...,x_{l_i}^i\right)$ is the bitstream *i*, l_i is the length of X_i. The bit frequency statistics mainly checks the bit frequency distribution of zero and one in a bitstream. Taking the bit frequency statistics parameter calculating process of X_i as an example, based on the $y_j^i=2x_j^i-1$ transformation, we change X_i to be a new sequence $Y_i=\left(y_1^i,y_2^i,...,y_{l_i}^i\right)$ composed of -1 and 1, and then get the binomial sum of the sequences as shown in formula (1).

$$S_i=y_1^i+y_2^i+...+y_{l_i}^i\,. \tag{1}$$

Further normalize the binomial sum of sequence as shown in formula (2).

$$F_{X_i}=\frac{|S_i|}{l_i}\,. \tag{2}$$

Then F_{X_i} is the bit frequency statistical parameter of X_i. From the definition of F_{X_i}, we can know that if the bits of X_i are all zero or one, the maximum value of F_{X_i} is one. Generally, F_{X_i} is normal distributed.

(2) Runs statistical parameter mining

Run is composed by successive zero or one bit; there are zero runs and one runs respectively with different lengths in X_i. We set z_{ij} as the frequency of zero run, e_{ij} as the frequency of the one run in X_i, where *j* is the length of the run, γ_0 as the longest lengths of zero run, γ_1 as the longest lengths of one run. On above definitions, we define the run statistical parameter as in formula (3).

$$R_{X_i}=\frac{\left|Var\left(e_{ij}\right)-Var\left(z_{ij}\right)\right|}{Var\left(e_{ij}\right)+Var\left(z_{ij}\right)}, \tag{3}$$

where

$$Var\left(e_{ij}\right)=\frac{1}{\gamma_1}\sum_{j=1}^{\gamma_1}\left(e_{ij}-\tilde{e}_i\right)^2, \tag{4}$$

$$Var\left(z_{ij}\right)=\frac{1}{\gamma_0}\sum_{j=1}^{\gamma_0}\left(z_{ij}-\tilde{z}_i\right)^2. \tag{5}$$

According to the definitions of z_{ij}, e_{ij}, γ_0 and γ_1, the binomial sum of sequence Y_i can be expressed as:

$$S_i=\sum_{j=1}^{\gamma_1}je_{ij}-\sum_{j=1}^{\gamma_0}jz_{ij} \tag{6}$$

and then F_{X_i} can be expressed as:

$$F_{X_i} = \frac{\left| \sum_{j=1}^{\gamma_1} je_{ij} - \sum_{j=1}^{\gamma_0} jz_{ij} \right|}{\sum_{j=1}^{\gamma_1} je_{ij} + \sum_{j=1}^{\gamma_0} jz_{ij}} . \tag{7}$$

Formula (2) and (7) show that there is no simple linear relationship between F_{X_i} and R_{X_i} .

(3) Bit frequency within a block statistics parameter mining

As described above, bit frequency within a block mainly focuses on the frequency distribution of zero and one in a block with a certain block length. In this situation, m is the block length, the bitstream X_i can be divided into $H_{im} = \left\lfloor \frac{l_i}{m} \right\rfloor$ blocks. π_{ij} is the bit one frequency of the block *j*.

$$\pi_{ij} = \sum_{k=1}^{m} x^{i}_{(j-1)m+k} . \tag{8}$$

When the block length is *m*, define B_{X_i} as the bit frequency within a block statistical parameter of X_i as shown in formula (9).

$$B_{X_i} = \frac{\sum_{j=1}^{H_{im}} (j\pi_{ij} - \tilde{\pi}_i)^2}{H_{im}\Phi_i} , \tag{9}$$

where $\tilde{\pi}_i = \frac{1}{H_{im}} \sum_{j=1}^{H_m} j\pi_{ij}$ and $\Phi_i = \max_{1 \le j \le H_{im}} \left((j\pi_{ij} - \tilde{\pi}_i)^2 \right)$.

(4) Optimal block length selection

Before we give the optimal block length selection algorithm, we firstly give the following definitions.

Definition 1: σ_k is the variance of π_{ji} for cluster C_k .

$$\sigma_k = \frac{1}{H_m N_k} \sum_{i=1}^{H_m} \sum_{j=1}^{N_k} \left(\pi_{ji} - \tilde{\pi}_{ki} \right)^2 , \tag{10}$$

where $\tilde{\pi}_{ki} = \frac{1}{N_k} \sum_{j=1}^{N_k} \pi_{ji}$ is the average value of π_{ji} for cluster C_k , N_k is the number of bitstreams included in C_k .

Definition 2: $\tilde{\sigma}$ is the average value of σ_k for the bitstreams sets $C = (C_1, C_2, ..., C_p)$.

$$\tilde{\sigma} = \frac{1}{p} \sum_{k=1}^{p} \frac{1}{H_m N_k} \sum_{i=1}^{H_m} \sum_{j=1}^{N_k} \left(\pi_{ji} - \tilde{\pi}_{ki} \right)^2 . \tag{11}$$

When $\tilde{\sigma}$ obtains the minimum value, we can confirm that the frequencies of bitstreams in each cluster have least differences as the block length is *m*.

Definition 3: σ is the variance of all $\tilde{\pi}_{ki}$ for the bitstreams sets $C = (C_1, C_2, ..., C_p)$.

$$\sigma = \frac{1}{H_m p} \sum_{i=1}^{H_m} \sum_{k=1}^{p} \left(\tilde{\pi}_{ki} - \frac{1}{p} \sum_{k=1}^{p} \tilde{\pi}_{ki} \right)^2 . \tag{12}$$

When σ obtains the maximum value, we can confirm that the frequencies of bitstreams in different clusters have greatest differences as the block length is *m*.

Definition 4: Q_m is the difference of σ and $\tilde{\sigma}$ for optimal block length selection.

$$Q_m = \sigma - \tilde{\sigma} . \tag{13}$$

The purpose of Q_m definition is to balance $\tilde{\sigma}$ and σ . The optimal block length should ensure $\tilde{\sigma}$ is as small as possible, but σ is as large as possible. So when we get the maximum Q_m , we take *m* as the optimal block length. Based on above definitions, the main steps of the optimal block length selection algorithm are as follows:

Step 1: Calculate the bit frequency statistical parameters, runs statistical parameters and bit frequency within a block statistical parameters for all the bitstreams respectively, m_0 is the initial block length, H_{km_0} is the minimum block number defined in formula (14).

$$H_{km_0} = \left\lfloor \frac{\min(l_1, l_2, ..., l_N)}{m_0} \right\rfloor . \tag{14}$$

Step 2: Using the *k*-means algorithm cluster the bitstreams into *p* clusters as $C = (C_1, C_2, ..., C_p)$

Step 3: Set $m = m_0$, confirm Q_{m_0} according to formula (10), (11), (12) and (13).

Step 4: Set $m = m + 1$, get the new block number H_{km} , and then get the new Q_m according to formula (10) (11) (12) and (13).

Step 5: if $m < m_{\max}$, repeat Step(4), get corresponding Q_m for different block length.

Step 6: Select the optimal block length m_{opt} according to formula (15).

$$m_{opt} = \underset{m_0 \le m \le m_{\max}}{\arg\max} \{Q_m\} . \tag{15}$$

3.2 UNKNOWN PROTOCOL BITSTREAMS CLUSTERING BASED ON THE *K*-MEANS ALGORITHM

Once we get the *F*, *R* and *B* characteristic parameters of bitstreams, the bitstreams will be clustered by the *k*-means algorithm. The initial clustering centers selection and optimal clustering number selection algorithms for the *k*-means algorithm are as follows:

(1) Initial clustering centers selection algorithm

(a) Confirm the range of characteristic parameters for each dimension as $[u_{j_{\min}}, u_{j_{\max}}]$, where $1 \le j \le h$ and h is the maximum dimension number.

(b) Set λ_1 as the number of sections for sample density statistics of the first dimension, $\varphi_1(i)$ is the sample density of section *i*.

$$\phi_1(i)=\frac{\Delta N_1(i)}{\Delta u_1},1\le i\le\lambda_1 . \tag{16}$$

(c) If $\varphi_1(m)$ is a peak value, the sample density statistics sections between $\varphi_1(m)$ and the previous candidate is not less than η_1, the average value of section *m* is the candidate for clustering center.

(d) Adjust corresponding parameters, until *h* is equal to the maximum dimension, then return the candidates for initial clustering centers.

(e) Initial clustering centers selection

Set up the relationship tree of the candidates for initial clustering centers according to their mapping relationships of each dimension. The initial clustering centers selection process is based on the MMD algorithm.

(2) λ_j and η_j selection

As the average sample density of each dimension may be different, the values of λ_j as defined in formula (17) should be also different. If $W_j > W_{j+1}$, there will be more sample density statistical sections in dimension *j* than dimension *j*+1. W_j is the difference of the maximum value and minimum value of dimension *j*. k , N and h usually satisfy $k << N$, $h << N$ and $\sqrt[h]{N}\ge k$. The parameter $\frac{W_j}{\sqrt[h]{\prod_{i=1}^{h}W_i}}$ is the section number inching parameter for the dimension *j*.

$$\lambda_j=\frac{W_j}{\sqrt[h]{\prod_{i=1}^{h}W_i}}\sqrt[h]{N} . \tag{17}$$

There may be many density peak values in the overlap sections among clusters. So when we check the peak values, the peak values in the η_j sections radius will be ignored. The parameter η_j is defined in formula (18) and the corresponding conclusions are as follows.

$$\eta_j=\frac{\lambda_j-k}{2k} . \tag{18}$$

Conclusion 1: If there is no overlap structure between any two clusters in the dimension *j*, η_j can make sure that the selected initial clustering centers are all included in their clusters.

In order to prove the conclusion 1, we give an example of clusters distributions as shown in Figure 1. To cluster 1, the maximum and minimum value in the direction of x are x_{12} and x_{11}, the maximum and minimum value in the direction of y are y_{12} and y_{11} . To cluster 2, the maximum and minimum value in the direction of x are x_{22} and x_{21}, the maximum and minimum value in the direction of y are y_{22} and y_{21} . Where $y_{12} > y_{22}$ and $y_{11} > y_{21}$, the proof of conclusion 1 is as follows.

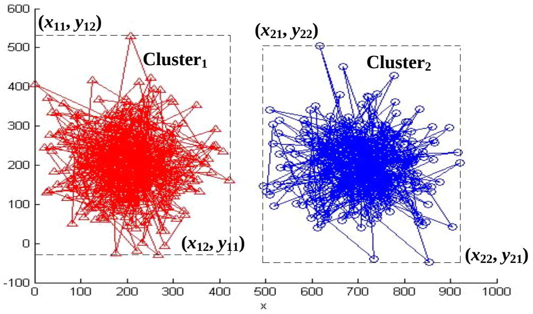


FIGURE 1 Example of clusters distributions

Proof: Based on the above definitions and according to formula (17) λ_x can be expressed as:

$$\lambda_x=(x_{22}-x_{11})\sqrt{\frac{N}{(x_{22}-x_{11})(y_{12}-y_{21})}} . \tag{19}$$

After we confirmed the λ_x, the length of singe sample density statistic section in the direction of x is:

$$\Delta u_x=\frac{W_x}{\lambda_x}=\sqrt{\frac{(x_{22}-x_{11})(y_{12}-y_{21})}{N}} . \tag{20}$$

According to formula (18), η_x can be obtained as follows:

$$\eta_x=\frac{\lambda_x-2}{4}=\frac{(x_{22}-x_{11})\sqrt{\frac{N}{(x_{22}-x_{11})(y_{12}-y_{21})}}-2}{4} . \tag{21}$$

The length of the sections which the parameter η_x corresponding is:

$$\Delta x=\eta_x\Delta u_x=\frac{(x_{22}-x_{11})-2\Delta u_x}{4} . \tag{22}$$

We can assume that the coordinates of the density peak values of Cluster1 and Cluster2 in the direction of x are O_{x1} and O_{x2}, where

$$\begin{cases} O_{x1}=\dfrac{x_{12}+x_{11}}{2} \\ O_{x2}=\dfrac{x_{22}+x_{21}}{2} \end{cases} . \tag{23}$$

$\overline{W}_x$ is the distance of O_{x1} and O_{x2} in the direction of x as shown in formula (24).

$$\overline{W}_x=O_{x2}-O_{x1} . \tag{24}$$

S_x is the difference of $\overline{W}_x$ and Δx as shown in formula (25).

$$S_x=\overline{W}_x-\Delta x . \tag{25}$$

If there is no overlap structure between Cluster1 and Cluster2 in the direction of x , where $x_{22} > x_{11}$ and $x_{21} > x_{12}$, S_x satisfies $S_x > 0$. Conclusion 1 holds.

Conclusion 2: When there are some overlap structures between any two clusters, η_j can also make sure that the selected initial clustering centers are all included in their clusters.

Proof: As shown in Figure 2, we can assume that the clustering centers of cluster1 and cluster2 are $\frac{x_{11}+x_{12}}{2}$ and $\frac{x_{21}+x_{22}}{2}$ in the direction of *x*, the distance of them is Δx. The length of overlap structure is $\frac{W_x+2\Delta u_x}{2}$. As the samples in a cluster are normal distributed around their cluster center, the density peaks generally appear in $\left[\frac{x_{11}+x_{12}}{2},\frac{x_{21}+x_{22}}{2}\right]$. We will respectively analyze the distributions of clustering centers according to the relationships of $\left|\frac{x_{21}+x_{22}}{2}-\frac{x_{11}+x_{12}}{2}\right|$ and Δx.

(a) If $\left|\frac{x_{21}+x_{22}}{2}-\frac{x_{11}+x_{12}}{2}\right| \le \Delta x$, there will be only an initial clustering center. Conclusion 2 holds.

(b) If $\left|\frac{x_{21}+x_{22}}{2}-\frac{x_{11}+x_{12}}{2}\right| > \Delta x$, the will be only a clustering center, two clustering centers or three initial clustering centers candidates. In summary, when there are some overlap structures between any two clusters in dimension *j*, the initial clustering centers are all included in their clusters. Conclusion 2 is proved.

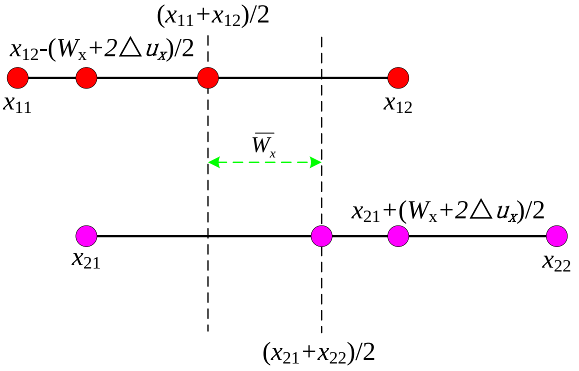


FIGURE 2 Example of clusters overlap structure

(3) The optimal clustering number selection

Once we selected the clustering algorithm, it is very important to establish an effective function $V\left(C^*\right)$ to evaluate the quality of clustering. As most of the current clustering validity functions are complex, based on sample density and clustering density we give a new clustering validity index called CVED (Clustering Validity Evaluation based on Density). When the division of the bitstreams is confirmed as $C^k=\left(C_1,C_2,...,C_k\right)$, $\tilde{\xi}$ is the sample density distribution of all samples of all dimensions, $\tilde{\rho}$ is the clustering density of all dimensions.

$$\tilde{\xi}=\frac{1}{hk}\sum_{j=1}^{h}\sum_{i=1}^{k}\frac{|C_i|}{W_{ij}}, \tag{26}$$

$$\tilde{\rho}=\frac{k}{h}\sum_{j=1}^{h}\frac{1}{\overline{W}_j}, \tag{27}$$

where $|C_i|$ is the number of the samples included in cluster C_i, W_{ij} is the difference of the maximum value and minimum value of dimension *j* for cluster C_i. $\overline{W}_j$ is the difference of the maximum and minimum clustering centers of dimension *j*.

Definition 6: $V\left(C^*\right)$ is the clustering validity index as shown in formula (28):

$$V\left(C^k\right)=\frac{\tilde{\xi}-\tilde{\rho}}{\tilde{\xi}+\tilde{\rho}}. \tag{28}$$

The optimal clustering number k_{opt} is confirmed according to formula (29):

$$k_{opt}=\underset{k_{\min}\le k\le k_{\max}}{\arg\max}\left\{V\left(C^k\right)\right\}. \tag{29}$$

3.3 ALGORITHM COMPLEXITY ANALYSIS

To facilitate the analysis, assuming the number of density sections in each dimension is λ. The operation times of density statistics for the first dimension are $2N\lambda$. The average ratio of effective density peak values is α, when we confirm the parameters of the dimension *j*+1 form the parameters of dimension *j*, the operation times are $\lambda\alpha\left(N\lambda+\lambda\right)$. If h is the number of the dimensions, the all operation times are $2N\lambda+\lambda\alpha\left(N\lambda+\lambda\right)+\cdots+(\lambda\alpha)^{h-1}\left(N\lambda+\lambda\right)$. When confirm the initial clustering centers by the MMD algorithm, the main complexity of the algorithm is to calculate the distances of the initial clustering centers, but the operation times of distances calculating can be ignored as the number of initial clustering centers is much smaller than the number of samples. So the complexity of our initial clustering centers selection algorithm is $O\left(N(\lambda\alpha)^{h-1}\right)$. In extreme cases, every sample density statistics section only contains a sample, where $\lambda^h \le N$ is satisfied according to formula (17), the actual complexity of the algorithm is far less than $O\left(N^2\right)$.

The main complexity of the proposed algorithm is mainly due to the process of clustering. When $k=k_{\max}$, the operations times for $\tilde{\xi}$ and $\tilde{\rho}$ are respectively $hk_{\max}$ and h. The complexity of the proposed algorithm is $O\left(hk_{\max}\right)$. The complexity of the CH, DB, KL and COPS indices is $O\left(N\right)$. The complexity of the Wint, IGP and BWP indices is $O\left(N^2\right)$.

4 Experimental results and analysis

4.1 EXPERIMENTAL SUBJECTS AND SETTINGS

In our experiment, the system of the computer is Windows XP, all bitstreams are from the internet including the HTTP, DNS, ICMP, TELNET and generic UDP bitstreams, the number of each dataset is different with each other. The detail information of HTTP, DNS, ICMP, TELNET and UDP bitstreams is shown in Table 1. In our experiments, we took the bitstreams of HTTP, DNS, ICMP, TELNET and UDP datasets as the bitstreams of unknown protocols.

TABLE 1 Data sets information

Datasets	Sample numbers	Dimensions
HTTP	285	3
DNS	47	3
ICMP	270	3
TELNET	102	3
UDP	1000	3

4.2 RESULTS AND ANALYSIS OF BITSTREAMS CHARACTERISTIC PARAMETERS SELECTION

To verify the affects of different block lengths to the character value of bit frequency within a block, we initially cluster the bitstreams of the HTTP, DNS, ICMP, TELNET and UDP datasets based on the *k*-means algorithm. When we calculate the characteristic value of bit frequency within a block, randomly choose 20 as the block length. In our experiment, the shortest length of the bitstreams is 320; we choose 160 as the longest block length, so the block length is ranging from 2 to 160, the values of Q_m for different block lengths are shown in Figure 3. As shown in Figure 3, the values of Q_m is ranging from -7.729 to 23.866, the maximum value of Q_m is 23.866 when the block length is 88. On the other hand, the minimum value of Q_m is -7.729 when $m \in [107,121]$. So when we calculate the characteristic value of bit frequency within a block, $m = 88$ is the optimal block length.

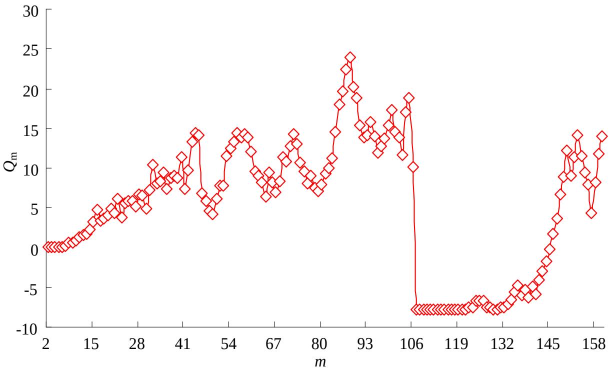
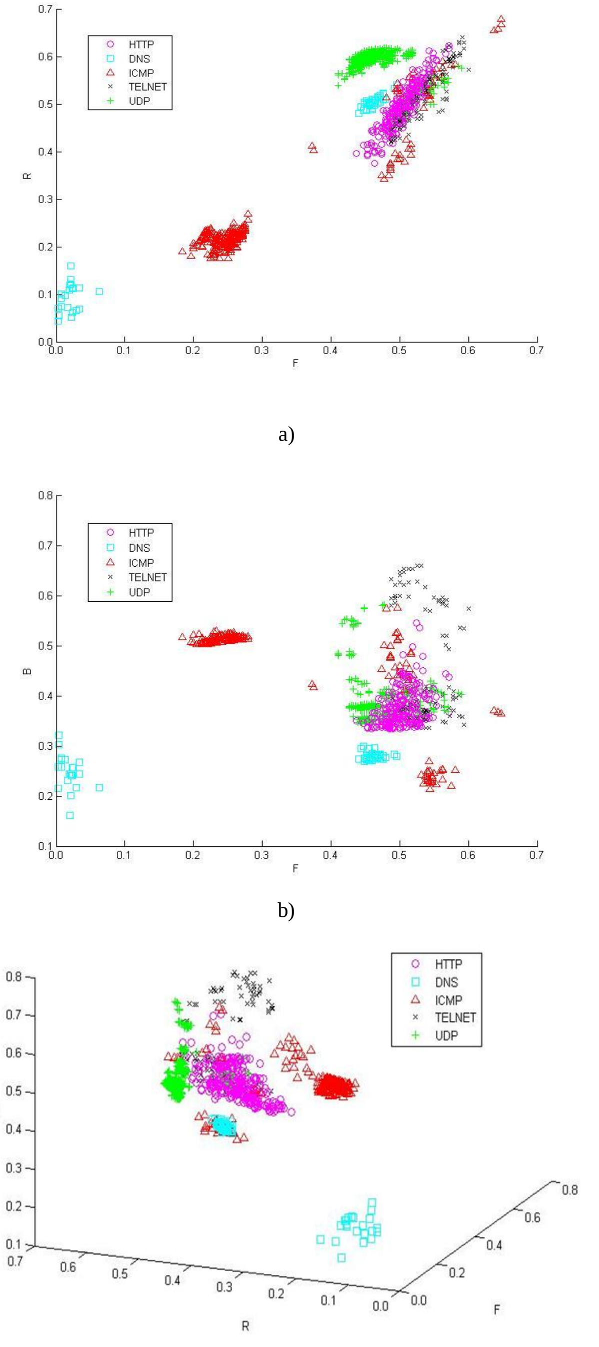


FIGURE 3 Q_m for different block length

When the block length is 88, the distributions of the *F* and *R* values are shown in Figure 4a), the maximum and minimum values of *F* are 0.6479 and 0.0026, the maximum and minimum values of *R* are 0.6788 and 0.0435. Meanwhile, the distributions of *F* and *B* are also shown in Figure 4b), the maximum and minimum values of *B* are 0.6601 and 0.1616. The three-dimensional distributions of *F*, *B* and *R* are shown in Figure 4c). Although there are some overlap structures among the characteristic parameters of the bitstreams, but most of the bitstreams have presented effective clustering characteristic, we can cluster them into corresponding bitstream datasets.



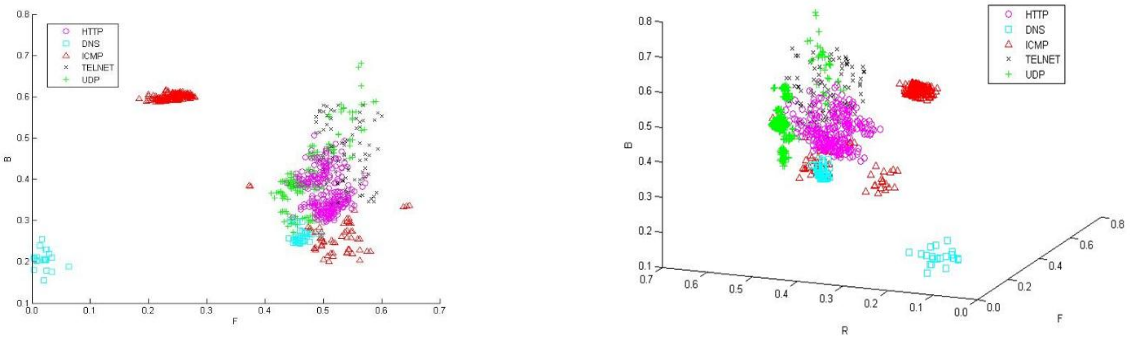
a)

b)

c)

FIGURE 4 Distributions of *F*, *R* and *B* for maximal Q_m: a) Distributions of *F* and *R*, b) Distributions of *F* and *B*, c) Distributions of *F*, R and *B*

To further illustrate the importance of selection optimal block length, under the conditions of $m = 120$ (the value of Q_m is minimum), we recalculate the *B* values for the bitstreams. The distributions of *F* and *B* are shown in Figure 5a). The distributions of *F*, *R* and *B* are shown in Figure 5b). In Figure 5a) and Figure 5b), there are more overlap structures of the *B* values. The clustering characteristic of the *B* values in Figure 5a) are absolutely more indistinctive than the *B* values in Figure 4b). The experimental results demonstrate the validity of the proposed optimal block length selection algorithm.



a)

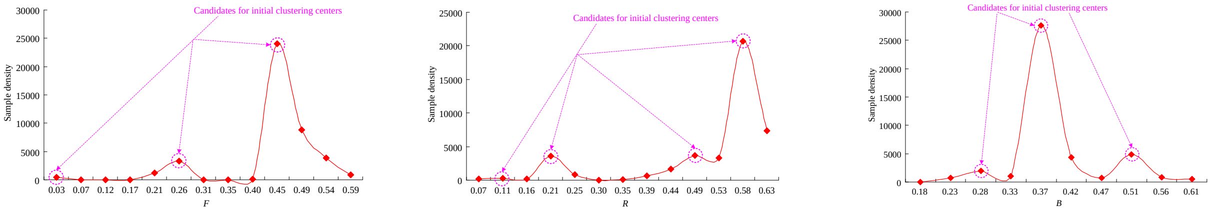
b)

FIGURE 5 Distributions of *F*, *R* and *B* for minimal Q_m : a) Distributions of *F* and *B*, b) Distributions of *F*, *R* and *B*

4.3 BITSTREAMS CLUSTERING RESULTS AND ANALYSIS

(1) Initial clustering centers selection

With the proposed algorithm, we can get the sample density distribution characteristics of the HTTP, DNS, ICMP, TELNET, and UDP bitstreams as shown in Figure 6a), Figure 6b) and Figure 6c). According to formula (17), we can respectively calculate the section numbers in the direction of *F*, *R* and *B*; they are 13, 10 and 13. The values of η_F , η_R and η_B can also be confirmed by formula (18), they are 0.8, 0.8 and 0.5. As η_F , η_R and η_B are all less than 1, so all of the peak values in the directions of *F*, *R* and *B* should all be taken as the candidates for initial clustering center. There are three candidates for initial clustering center both in the direction of *F* and *B*, their coordinates are F_1=0.03, F_2=0.26, F_3=0.45, B_1=0.28, B_2=0.37 and B_3=0.51. In the direction of *R*, there are four candidates for initial clustering center, their coordinates are R_1=0.11, R_2=0.21, R_3=0.49 and R_4=0.58. Based on the MMD algorithm and the relationship tree of the candidates for initial clustering center, we obtained five initial clustering centers, they are (0.03,0.11,0.28), (0.26,0.21,0.51), (0.45,0.49,0.28), (0.45,0.58,0.37), (0.51,0.49,0.37).



a)

b)

c)

FIGURE 6 Distributions of sample density: a) Sample density in the direction of *F*, b) Sample density in the direction of *R*, c) Sample density in the direction of *B*

(2) Similarity analysis of clustering centers and impacts on the iteration times

To illustrate the effectiveness of the proposed initial clustering centers selection algorithm, the similarity value of the initial clustering centers $U'=(u'_1,u'_2,...,u'_k)$ and the final clustering centers $U=(u_1,u_2,...,u_k)$ is defined in formula (30).

$$\tau_i = \frac{4(u_i,u'_i)}{\left(|u_i|+|u'_i|\right)^2} . \qquad (30)$$

When the *k*-means algorithm is converged, we get the final clustering centers, they are (0.02, 0.09, 0.24), (0.25, 0.22, 0.51), (0.50, 0.50, 0.36) and (0.46, 0.59, 0.39). The similarity values of the initial clustering centers and final clustering centers are 99.47 %, 99.97%, 99.40%, 99.98% and 97.36%. Furth more, we also get the average similarity values of the initial clustering centers and the final clustering centers by respectively running the RS, MMD and our initial clustering centers selection algorithm. The results are shown in Figure 7. During 100 repeated experiments, the constant average similarity value of our algorithm is 99.24%. As shown in Figure 7, the average similarity values of the RS method are unstable due to the randomness of the clustering centers; its value is ranging from 86.25% to 99.80%. On the other hand, the average similarity values of the MMD algorithm are less unstable than the RS method as there is only one random clustering center; its value is ranging from 91.88% to 98.63%.

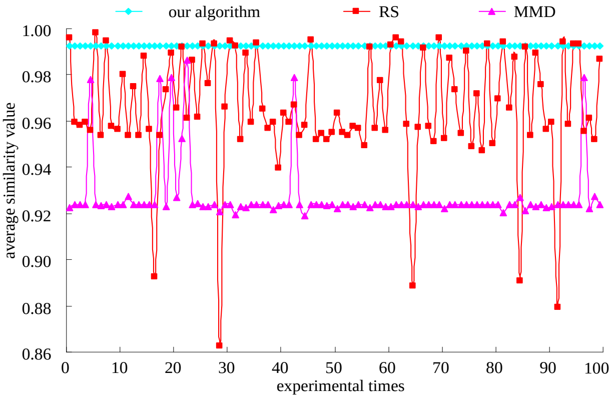


FIGURE 7 Average similarity values

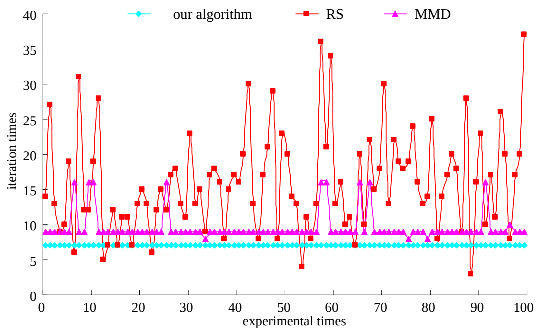
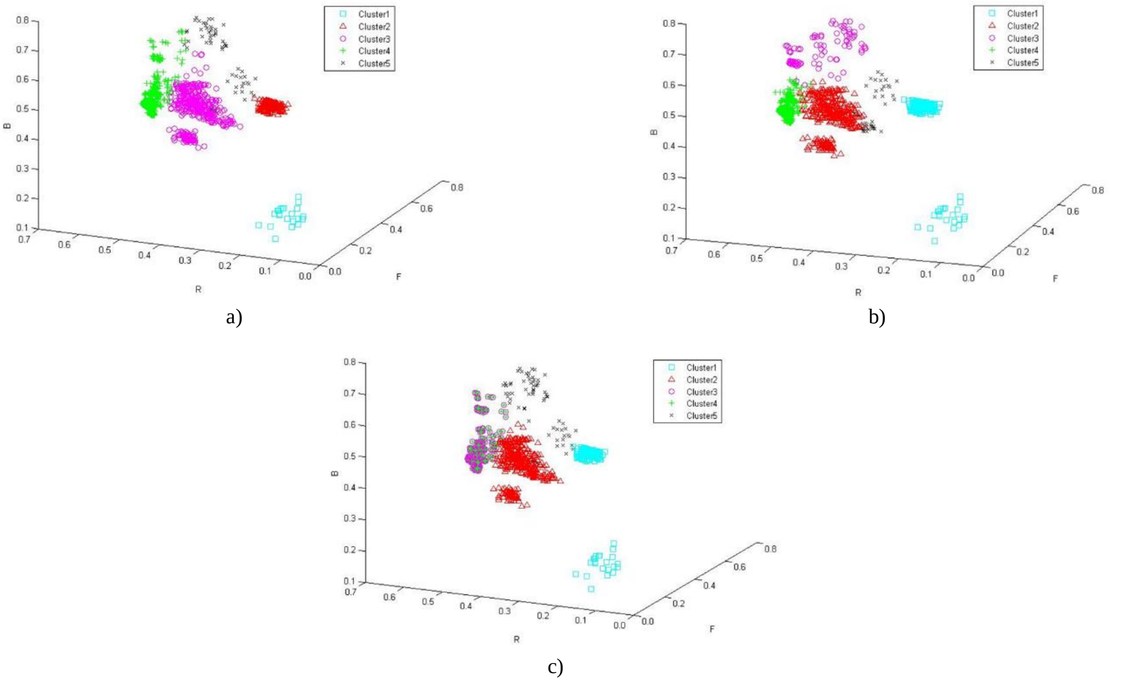


FIGURE 8 Iteration times

To verify the effect of the initial clustering centers to the iteration times of the *k*-means algorithm, we run the RS method, MMD algorithm, our initial clustering centers selection algorithm and the *k*-means algorithm for 100 times. The iteration times of the *k*-means algorithm are shown in Figure 8. As shown in Figure 8, when using our algorithm, the iteration times of the *k*-means algorithm is 7, but to the RS method and MMD algorithm the iteration times of the *k* -means algorithm are respectively ranging from 3 to 37, 8 to 16. Although, the iteration times 3 from the RS method is less than 7 from our algorithm, the clustering results of our algorithm are steadier than the RS method and MMD algorithm.

(4) Impacts on the clustering results

To verify the effect of the initial clustering centers to cluster results, we set 5 as the number of the initial clustering centers, the clustering results of the *k*-means algorithm for our algorithm, RS method and MMD algorithm are respectively shown in Figure 9a), Figure 9b) and Figure 9c). The results of our algorithm are more close to the original clustering characteristics of bitstreams in Figure 4c).



a)

b)

c)

FIGURE 9 Affects of initial clustering centers to clustering results: a) Clustering results of our algorithm, b) Clustering results of the RS method, c) Clustering results of the MMD algorithm

(5) Optimal clustering number selection

In order to verify the effectiveness of the CVED index, we have calculated the values of the KL, Wint, IGP, COPS, BWP and CVED indices and given the clustering number of these indices referring to as shown in Table 2. The experimental numbers of clusters from the KL, Wint, IGP and COPS indices are larger than the actual number of clusters due to the dispersive distributions of the bitstreams. On the other hand, the experimental numbers of clusters from the BWP and CVED indices are closer to the actual number of clusters.

TABLE 2 Numbers of clusters for different indices

Indices	Actual values	Experimental values
KL	5	9
Wint	5	10
IGP	5	7
COPS	5	8
BWP	5	6
CVED	5	6

5 Conclusions

In order to get the characteristic parameters of the bitstreams from the aspect of independent protocol, we defined the characteristic parameters of bit frequency, runs and bit frequency within a block for bitstream respectively. As the characteristic parameter of bit frequency within a block is sensitive to the block length, we proposed an algorithm based on the principle of the variance to obtain the optimal block length. As the sample density in each cluster is generally higher than the average sample density, we firstly calculated the sample density in each sample density calculating section for every dimension, the average sample value of section with the density peak value is taken as the candidate for initial clustering center. The relationship tree of candidates for initial clustering centers is set up based on the mapping relationships of dimensions. With the combination of the MMD algorithm, the initial clustering centers are selected from the relationship tree.

Furthermore, we also defined the function of clustering quality evaluation based on the definitions of sample density in cluster and cluster density. Taken the bitstreams of HTTP, DNS, ICMP, TELNET and UDP datasets as unknown protocol bitstreams, the experimental results demonstrate that our proposed algorithms can effectively mine the characters of protocol bitstreams and divide the bitstreams into the corresponding clusters. However, with the considerations of multi-value property of protocol field, there are some overlap structures among F, R and B values respectively which have some affects to the bitstreams clustering. Our next research work is to mine more effective parameters for unknown protocol bitstreams.

Acknowledgments

This work is supported by the National Natural Science Foundation of China. The authors gratefully acknowledge assistance and encouragement from Prof. Yuri Shunin. The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

References

[1] Zheng T M, Wang T, Guo S Z 2012 Improved space protocol identification algorithm *Journal on Communications* **33**(5) 183-90

[2] Tan J, Chen X S, Du M 2012 A novel real-time p2p identification algorithm based on BPSO and neural networks *Journal of Central South University (Science and Technology)* **43**(6) 2190-97

[3] Chen L F 2008 Research on clustering methods for high dimensional data and their applications *PhD dissertation of Xia Men University*.

[4] Charles V W, Fabian M, Gerald M M 2006 On inferring application protocol behaviors in encrypted network traffic *Journal of Machine Learning Research* **7**(12) 2745-69

[5] Sun G L, Xue Y B, Dong Y F 2010 A novel hybrid method for effectively classifying encrypted traffic *GLOBECOM 2010*: *Proc. Communications and Systems Security (Miami, Florida, USA, 6-10 December 2010) IEEE* 2010, pp 1-5

[6] Bo Z 2012 Research on protocol independent online identification of encrypted traffic *PhD dissertation of PLA Information Engineering University*

[7] Zhao B, Gong H, Liu Q R 2013 Protocol independent identification of encrypted traffic based on weighted cumulative sum test *Journal of Software* **24**(6) 1334-45

[8] Xu R 2005 Survey of clustering algorithm *IEEE Tran on Neural Networks* **16**(3) 645-78

[9] Zhou S B 2011 Research and application on determining optimal number of clusters in cluster analysis *PhD dissertation of Jiang Nan University*

[10] Likas A, Ulassis M, Uerbeek J 2003 The global k-means clustering algorithm *Pattern Recognition* **36**(2) 451-61

[11] Liu Y M, Zhang H X 2011 Approach to selection initial centers for *k*-means with variable threshold *Computer engineering and applications* **47**(32) 56-8

[12] Xiong Z Y, Chen R T, Zhang Y F 2011 Effctive method for cluster' initialization in K-means clustering. *Application Research of Computers* **28**(11) 4188-90

[13] Li X, Zhang J F, Cai J H 2012 A fuzzy clustering algorithm based on large density region *Journal of Chinese Computer Systems* **33**(6) 1310-5

[14] Chen G P, Wang W P, Huang J 2012 Improved initial clustering center selection method for *k*-means algorithm *Journal of Chinese Computer Systems* **33**(6) 1320-3

[15] Calinski T, Harabasz J 1974 A dendrite method for cluster analysis *Communications in Statistics* **3**(1) 1-27

[16] Davies D L, Bouldin D W 1979 A cluster separation measure *IEEE Transactions on Pattern Analysis and Machine Intelligence* **1**(2) 224-7

[17] Dudoit S, Fridlyand J 2002 A prediction-based resampling method for estimating the number of clusters in a dataset *Genome Biology* **3**(7) 1-21

[18] Dimitriadou E, Dolnicar S, Weingessel A 2002 An examination of indexes for determining the number of cluster in binary data sets *Psychometrika* **67**(1) 137-60

[19] Kapp A V, Tibshirani R 2007 Are clusters found in one dataset present in another dataset? *Biostatistics* **8**(1) 9-31

Authors	
	Wu Yang, 1985, Cheng Du, China **Current position, grades:** PhD. Student **University studies:** Shijiazhuang Mechanical Engineering College **Scientific interest:** Network security and network data analysis **Experience:** From 2005 to 2009, studied in UESTC and got the Bachelor Degree in 2009; From 2009 to 2012, studied in Shijiazhuang Mechanical Engineering College and got the Master Degree in 2012; From 2012 to now, studies in Shijiazhuang Mecha nical Engineering College for Doctor Degree.
	Wang Tao, 1964, Shi Jia-zhuang, China **Current position, grades:** Professor **University works:** Shijiazhuang Mechanical Engineering College **Scientific interest:** network security and cryptology **Publications:** Principles and Methodologies of Side-Channel Analysis in Cryptography, published by the Science Press, 2014. **Experience:** From 1990 to now, works in the Department of Information Engineering of Shijiazhuang Mechanical Engineering College.
	Li Jin-dong, 1990, Shi He-zi, China **Current position, grades:** Graduate Student **University studies:** Shijiazhuang Mechanical Engineering College **Scientific interest:** Network security and network data analysis **Experience:** From 2008 to 2012, studied in Xinjiang University and got the Bachelor Degree in 2012; From 2009 to now, studies in Shijiazhuang Mechanical Engineering College for the Master Degree.