Network Attack Classification and Recognition Using HMM and Improved Evidence Theory

Gang Luo
College of Computer Science and Electronic Engineering
Hunan University
Changsha, China

Ya Wen
College of Computer Science and Electronic Engineering
Hunan University
Changsha, China

Lingyun Xiang
Hunan Provincial Key Laboratory of Intelligent Processing of Big Data on Transportation
Changsha University of Science and Technology
Changsha, China

***Abstract*—In this paper, a decision model of fusion classification based on HMM-DS is proposed, and the training and recognition methods of the model are given. As the pure HMM classifier can't have an ideal balance between each model with a strong ability to identify its target and the maximum difference between models. So in this paper, the results of HMM are integrated into the DS framework, and HMM provides state probabilities for DS. The output of each hidden Markov model is used as a body of evidence. The improved evidence theory method is proposed to fuse the results and encounter drawbacks of the pure HMM for improving classification accuracy of the system. We compare our approach with the traditional evidence theory method, other representative improved DS methods, pure HMM method and common classification methods. The experimental results show that our proposed method has a significant practical effect in improving the training process of network attack classification with high accuracy.**

***Keywords*—Hidden Markov Model; Evidence theory; Network attack; KDD CUP99; Classification**

I. INTRODUCTION

With the development and popularity of Internet, the network environment in today's society is more and more complex. Security of network has become a very important problem in the network. Intrusion detection system which attempts to use data mining and machine learning methods to detect and classify intrusion activities plays an important role in detecting and preventing network attacks[1]. However, intrusion detection systems can be split into two groups: 1) anomaly-based detection system and 2) misuse-based detection system[2]. Each of them has a different way in detecting and protecting data security and has both advantages and disadvantages. The misuse-based detection system, especially the reasoning system based on model matching, can achieve high classification accuracy for known attacks. Scholars proposed various classifier models to solve classification problem in network intrusion detection, including Bayesian network, fuzzy logic, k-nearest neighbor, decision tree, neural networks, support vector machine, the hidden Markov model.

Cheng Xiang [3] proposed a multiple-level hybrid classifier, a novel intrusion detection system, which combined supervised tree classifiers and unsupervised Bayesian clustering to detect intrusions. The performance of this approach was shown to have high detection and low false alarm rates. In [4], a multiple classifier intrusion detection model was presented, which was based on a new data mining method called hidden Naive Bayesian. This method was better than other models, but it only had a high detection rate for the DoS (the denial of service) attack while the other attack detection accuracy was not high. Yuk [5] applied intelligent dynamic swarm based rough set for feature selection and simplified swarm optimization for intrusion data classification. The performance of the hybrid intrusion detection system on KDD Cup 99 dataset is better than others with high classification accuracy.

Some researchers use machine learning methods to design intrusion detection systems. Most of them are based on SVM technology which has a solid theoretical basis and can classify data records into multiple classes. Horng et al. [6] proposed an SVM-based intrusion detection system based on a hierarchical clustering algorithm to preprocess dataset before training. The simple feature selection procedure was applied to eliminate unimportant features from the training set so that the obtained SVM model could classify the network traffic data more accurately. However, this system showed better performance in the detection of DoS and Probe attacks but not very good in U2R and R2L attacks. In [7], a pipeline of the data preprocess and data mining was put forward in IDS to choose critical features. With the combination of clustering method and support vector machine, an efficient and reliable classifier was developed to judge a network. The performance of SVM was good in data classification, but not suitable for large scale dataset. Training complexity is deeply dependent on the data volume of training dataset, and the greater amount of data will lead to higher training complexity. However, many data mining applications involve millions or even billions of pieces of data records. The system failure caused by the lack of memory makes the SVM can't run such a large dataset.

The Markov model and hidden Markov model which are initially used for speech recognition (Rabiner,1989), handwriting recognition (Gunter and Bunke, 2003), biological sequence analysis (Durbin et al., 2006) have been applied to computer security model in recent years. In the field of computer security, HMM is mainly used for anomaly-based

intrusion detection systems. Warrender et al. (1999) made a pioneering work in this area and they use HMM for system modeling. HMM can also be used in network security. Ariu [8] proposed a novel solution where the HTTP payload is analyzed using hidden Markov model. The proposed system, named HMMPayl, had high classification accuracy and was very effective on most common attacks of the Web application. The core idea of attack classification is pattern recognition. Hidden Markov model can effectively describe the hidden Markov process containing unknown parameters. HMM can get hidden parameters of the process from observable feature parameters, and use these parameters to make further analysis for attack classification [9]. However, when the dimension of feature parameters space is high, the training structure is complex, the training time is very long, and the classification recognition accuracy is quite low. In intrusion detection, several attacks may show some similar features. That is, under certain features, the attacks are likely to have a certain probability of occurrence. Fusion all kinds of feature information to obtain the occurrence probability of each attack and the maximum probability of occurrence can be judged as the main attack. Therefore, the use of evidence theory [10-11] is particularly suitable for classification and recognition of information fusion.

The study of neural biology showed that the information process of biological sensing system can be divided into two relatively independent processing procedures: information unit decomposition and fusion. Such way of early decomposition and late fusion with high ability in information processing and intelligent decision [12]. According to this, the idea of multi-features fusion and decision making can be used in the classification of attacks to achieve the purpose of improving classification accuracy. Assume that there is a classification problem of N kinds of attacks. The whole attack feature parameters space is divided into K subspaces according to certain rules. Then decision model of each feature subspace is constructed to achieve the mapping of feature subspace. If use hidden Markov model, K feature subspace will lead to K hidden Markov models process, and K diagnostic results will be obtained. This process is equivalent to the decomposition of information unit. K diagnostic results of K sub hidden Markov models are then used as K bodies of evidence. By using the evidence theory to combine K bodies of evidence, the fusion of information units can be realized, and the final decision can also be made.

Therefore, in this paper, the hidden Markov model and the evidence fusion theory are applied to network security. A new information fusion system based on HMM and DS evidence theory is proposed which can effectively achieve the target of network attack classification and recognition. At the same time, a new method of evidence fusion based on entropy weight is proposed. By calculating the information entropy of source data to obtain weight of evidence, and modify the basic probability assign (BPA) of original evidence. Finally, the rule of combination is used to combine the modified BPA.

The rest of the paper is organized as follows: Section II briefly describes the principal theory of hidden Markov model and DS evidence theory. Section III presents the improved DS evidence theory and explains the details of the theoretical concept of the proposed HMM-DS system. The analysis of experimental results for KDD CUP99 using HMM-DS has been compared with C4.5, LibSVM, Naïve Bayes, which are presented in section IV. In Section V, the study concludes with a summary of the research undertaken.

II. Related Work

A. Hidden Markov Model

A hidden Markov model is a statistical model which is used to describe a Markov process containing unknown parameters. It is mainly used to determine the hidden parameters of the process from the observable parameters, and then use these parameters to make a further analysis. Figure 1 shows the general architecture of an instantiated HMM. The random variable x_i is the hidden state, y_i is possible observation, a_{ij} is transition probability matrix, and b_{ij} is emission probability matrix.

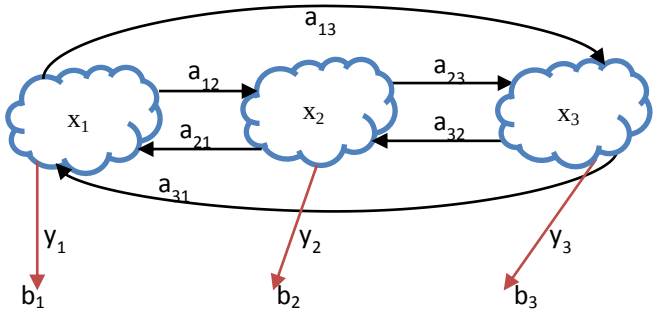


Fig. 1. The architecture of an instantiated HMM

An HMM can be described as five elements, which is $\{Q, O, \pi, A, B\}$.Q is the number of hidden states which is accurately known or guessed. O represents the number of observable states which can be achieved by training datasets. A is the matrix of state transition probabilities, B is probability distribution in each of the states which is also called the mixture matrix and π is the initial state of probability distribution. In state transition matrix and mixture matrix, each probability is independent of time. Namely, when the system is in an evolution, these matrices do not change over time. Therefore, we can use the compact notation $\lambda = \{\pi, A, B\}$ to denote an HMM with discrete probability. HMM can solve three problems [13]:

1) Evaluation problem. For a large number of sequences of HMMs (λ_1, λ_2, $\lambda_3 \ldots \lambda_k$) and observation sequence $O = \{o_1, o_2, \ldots, o_T\}$, Forward algorithm is used to calculate the probability of a given observation sequence, and then an HMM is chosen that best matches the observations.

2) Decoding problem. For a given model λ and observation sequence O, Viterbi algorithm is used to calculate the most likely sequence of hidden state.

3) Learning problem. For a given observation sequence and the related set of hidden states, Baum-Welch or Forward-Backward algorithm is applied for parameter estimation.

B. DS evidence theory

The evidence theory was first put forward by Dempster and developed by Shafer. In evidence theory, elements in the frame of discernment Θ are exclusive and exhaustive. Define m: $2^{\Theta}\rightarrow$ [0, 1] as basic probability assignment (BPA, also called mass function) satisfying: $\forall A\subset\Theta$, $m(\Phi)=0$, $\sum_{A\subset\theta} m(A)=1$

where A is called the focal element [14].

The core of DS evidence theory is the rule of combination. Two mass function m_1 and m_2, based on the evidence of two independent and reliable sources, can be combined into a new mass function by the use of conjunctive combination.

$$m_1 \oplus m_2(A)=\frac{1}{1-k}\sum_{B\cap C=A} m_1(B)m_2(C) \quad (1)$$

Where $k=\sum_{B\cap C=\varnothing} m_1(B)m_2(C), k\neq 1$ measures the conflict between m_1 and m_2. K is called the conflict coefficient. Dempster's rule of combination satisfies the associative law and the commutative law. There are n mass function (m_1, m_2,...m_n) in the frame of discernment, the conjunctive combination is calculated as

$$(m_1 \oplus \cdots m_n)(A)=\frac{1}{1-k}\sum_{\bigcap_{i=1}^{n} A_i=A} m_1(A_1)\cdots m_n(A_n) \quad (2)$$

Where $k=\sum_{\bigcap_{i=1}^{n} A_i=\varnothing} m_1(A_1)m_2(A_2)\cdots m_n(A_n)$.

As a kind of uncertain reasoning method, DS has attracted more and more attention. DS evidence theory can not only solve the problem of unknown and uncertainty, but also provides a very useful rule of combination which can help us fuse the evidence provided by multiple sources of evidence. Two key problems need to be solved in the process of DS evidence theory applied in data fusion classification. On the one hand, how to construct the basic probability assignment function of DS evidence theory, which is an important issue that must be solved in the process of combination and is not easy to determine. On the other hand, when the bodies of evidence to be combined are highly conflicting, counter-intuitive results may be obtained based on Dempster's rule of combination.

Scholars have proposed a variety of solutions to solve the issue. Some of them think that counter-intuitive results are caused by Dempster's rule, so they modified Dempster's rule to build a new combination rule. Yager [16] proposed an algorithm to distribute conflict belief to unknown proposition completely. This algorithm is more reasonable than that of D-S evidence theory in dealing with the combination paradox. However, the combination results are undesirable in combining multiple sources of evidence. In [17], Yager proposed a very interesting approach which made use of a weighted aggregation of the belief structures where the weights were related to the degree of dependence. It is too theoretical to be used in real applications. However, how to define the degree of dependence is not given. Sun [18] allocated part of the basic probability assignment of the conflict to the set of propositions supported by the evidences by a certain proportion. The difference between Yager and Sun is the proportion of the conflicts allocated. Thierry [19] presented a modified combination rule with mass function of dependent information sources. This rule used a special description of the body of evidence to ensure the combination, but the results given are very strange, and it does not consider the degree of confidence in the source of evidence. Destercke and co-workers [20] generalized the minimum rule of possibility theory, but did not respect the fundamental equivalence between belief functions and their empty set.

Some researchers deal with conflict evidences based on the method of modifying evidence source while keeping the combination rule unchanged. Haenni [21] thought it may not be the problem of combination rule when results were not matched with the real situation. However, the evidence of conflict should be modified. The rule of combination proposed by Murphy [22] was just to average all the BPAs of relevant hypothesis to get new belief assignments. This method can get good convergence effect, but the weight of each sensor in practical system is not the same. Yong Deng [23] put forward a novel sequence weighted evidence combination approach by using the variances of BOE sequences to generate the weights. In [24], the proposed method used training data to build a normal distribution model for each attribute of the data. Then, a nested structure BPA function was constructed by using the relationship between the test data and normal distribution model. To deal with the outer dependence, Su [25] proposed a model based on the intersection of influencing factors identified during the information propagating and evaluating process. The relative weights of BPAs for a specific element in the outer dependence phase and the relative weights of elements in the inner dependence phase were used as the discount coefficient in the discounting operator.

III. Decision Model of Information Fusion Classification Based on HMM-DS

A. Improved DS evidence theory

In this paper, the feature space is divided into K subspaces according to the character of the feature space, and the function of each subspace is different. Some features of KDD CUP99 may be irrelevant and some others may be redundant. The importance of each feature subspace is also different. The result is more accurate by obtaining the importance coefficient of features to get new basic probability assignment. In this paper, based on the information entropy of the specific features of data source, the entropy weight to determine the importance coefficient as the weight of the fusion feature is obtained.

The basic idea of entropy method is depended on the variability of indices to determine the objective weight. The smaller the information entropy of the index is, the greater the degree of variation of the index value, and the more information it will provide. In the comprehensive evaluation, the index can play a bigger role with high weight. On the contrary, the greater the information entropy of the index is, the smaller the degree of variation of the index value, and the less information will be provided. Then the index just plays a smaller role with low weight [26].

The following procedure has been used in building the weight vector.

1) Assume that all the features of the source data are: $(X_1,X_2,X_3,\ldots,X_n)$, and $X_j=\{x_{1j}, x_{2j},\ldots x_{mj}\}$,which represents m record of evaluating data with n features. Equation (9) is used for data standardization and a data matrix $R=(r_{ij})_{m\times n}$ will be get after standardization of all indexes.

$R=\begin{Bmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ r_{21} & r_{22} & \cdots & r_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ r_{m1} & r_{m1} & \cdots & r_{mn} \end{Bmatrix}$, r_{ij} represents the data value of j *-th* index of the i *-th* record.

2) Calculate the proportion of the index value of the j *-th* index of the i *-th* item $p_{ij}=r_{ij}/\sum_{i=1}^{m} r_{ij}$. If $p_{ij}=0$, define $\lim_{p_{ij}=0} p_{ij}\ln p_{ij}=0$.

3) The information entropy of the j *-th* feature index is:

$$E_j=-\ln(m)^{-1}\sum_{i=1}^{m} p_{ij}\ln p_{ij} \tag{3}$$

4) According to the calculation formula of information entropy, the information entropy of each index is calculated like $S_i=(E_1,E_2,\ldots Ep), i=1,2,\ldots k$. The sum of the indexes of feature subspace is calculated by $H_i=\sum_{j=1}^{p} E_j$, and weight coefficient of H_i obtained as follows:

$$w_i=\frac{1-H_i}{k-\sum_{i=1}^{k} H_i} \quad i=1,2,\cdots,k \tag{4}$$

5) Based on weight coefficients of each evidence, weight vector can be obtained. $W=(w_1,w_2,\cdots,w_k)$. The basic probability assignment $m_i(A_j)$ of each element in the frame of discernment was modified by the weight vector.

$$m_i^*\left(A_j\right)=w_i * m_i\left(A_j\right) \tag{5}$$

6) In the equation(5), $j=1,2,\ldots,n$, n is the number of focal elements in the frame of discernment except for Φ. But the sum of basic probability assignment $m_i^*(A_j)$ value after adjustment is not 1 which does not meet the requirements of basic probability assignment function definition. In order to satisfy the definition of basic probability assignment, a definition is in need.

$$m_i^*\left(\Theta\right)=1-\sum_{j=1}^{n} m_i^*\left(A_j\right) \tag{6}$$

The basic probability distribution function is defined by (5) and (6). Finally, the combination in (2) is used to combine the modified evidences.

B. HMM-DS System Design

The whole feature parameter space is divided into several sub parameter spaces, and then a HMM model is designed for each feature parameter subspace. Parameters of each HMM model will be constructed with training data. After that, the models have the ability to learn. Meanwhile, these sub-models can form the preliminary judgment layer. Outcome probabilities based on every sub-HMM model can be obtained and these probabilities will act as basic probability assign of evidences in the frame of discernment. In the meantime, considering the problem of consistency variation among evidences, the improved evidence theory is used to fuse them to get the result of the cooperation of each sub model, and to improve the recognition accuracy of attack classification.

The combination of HMM and DS evidence theory can have complementary advantages, and it is beneficial to improve the speed and accuracy of classification identification. In this paper, the HMM-DS attack fusion classification decision model is shown in Figure 2. In this model, the whole classification process is divided into two layers: a preliminary identification layer based on HMM; a fusion decision layer based on HMM-DS evidence theory.

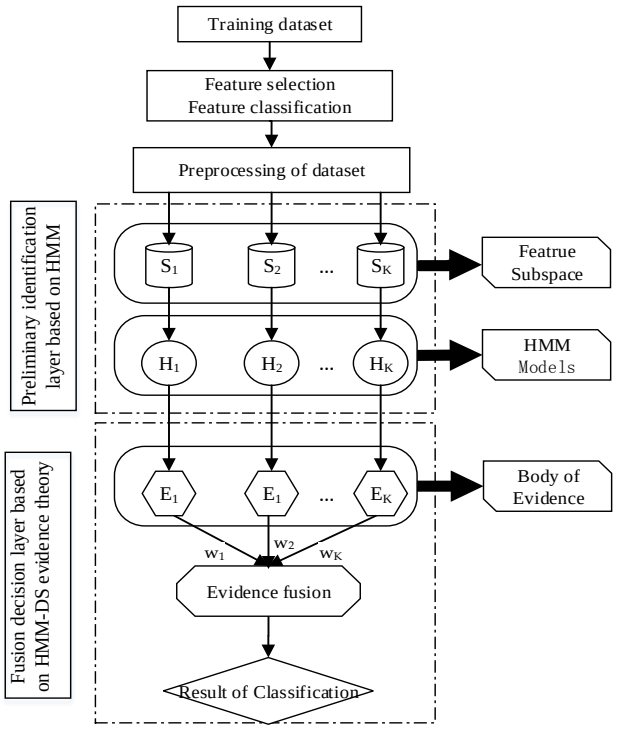


Fig. 2. HMM-DS system design

1) Preliminary identification layer based on HMM

Firstly, the feature parameter space S is divided into K sub-parameter space S_i, According to the definition of the parameter space, the corresponding learning dataset of sub HMM is obtained. Secondly, each independent sub hidden Markov model H_i is constructed and trained with learning dataset, and is capable of learning. Finally, the corresponding test samples are used to test the trained hidden Markov model, and the results obtained are the basis for the fusion decision layer in the next step.

2) Fusion decision layer based on HMM-DS theory

In the fusion decision layer, the output of each sub model in the preliminary recognition layer is used as a body of evidences E_i. The improved evidence combination method is used to fuse the evidence and obtain the final decision results

to achieve the attack classification. The creative process consists of the following five steps:

a) Establish a framework of discernment. According to expert experience and previous history records, establish an identification framework of discernment, $\Theta = \{A_1, A_2, \ldots, A_M\}$. In this paper, the proposition in the framework of recognition is corresponding to the attack type: Normal, Probe, Dos, U2R, R2L.

b) Construction of evidence. The output of each sub-HMM model is used as a body of evidence.

c) Calculate BPA of every element in the frame of discernment. DS evidence theory does not give the general calculation method of basic probability assignment, and methods used in relative papers were also different. The evaluation problem of HMM can obtain probability according to the observation sequence. Therefore, the BPA can be directly obtained by the probability calculated by the forward algorithm. As DS requires the sum of BPA of all elements must be 1, the probability gotten from HMM need to be normalized. The evidence H_i assigns to the BPA of proposition A_j can be expressed as follows:

$$m_i(A_j) = \frac{H_i(A_j)}{\sum_{j=1}^{M} A_j} \quad M \text{ is the number of attack type} \qquad (7)$$

Original evidences are modified by the improved evidence method in this paper, and new $m_i^*(A_j)$, $m_i^*(\Theta)$ is acquired

d) Evidence combination. DS evidence fusion method can be used to calculate the fused BPA $m(A_j)$.

e) Decision making. Decision methods used in evidence theory includes: decision making based on belief function, decision making based on minimum risk and decision making based on basic probability assignment. In this paper, the third method was used. That is, if $A_1, A_2 \subset U$, satisfy $m(A_1) = max\{m(A_i), A_i \subset U\}$; $m(A_2) = max\{m(A_i), A_i \subset U \text{ and } A_i \neq A_1\}$.

$$\text{If} \quad \begin{Bmatrix} m(A_1) - m(A_2) > \varepsilon_1 \\ m(\Theta) < \varepsilon_2 \\ m(A_1) > m(\Theta) \end{Bmatrix} \qquad (8)$$

Where A_1 is the result of the decision. Among them, ε_1 and ε_2 are the predefined thresholds. Θ is the uncertainty set.

IV. Experimental results and analysis

The experiments run in an Intel Pentium 2.7 GHz computer with 2.0G memory running Windows7.The code for data processing and data mining is written in MatlabR2014a.

A. KDD Cup99 Dataset description

The experiment data used in this paper is a benchmark database downloaded from KDD Cup99 [27]. This dataset contained a wide variety of intrusion simulated in a military network environment. It consists of two dataset, the training dataset and test dataset. Each network connection record is marked as Normal or Attack. The classification of attack behavior is a 5- class problem, and each network connection belongs to one of the following behavior: normal, denial of service (DOS), unauthorized access from a remote machine (U2R), unauthorized access to local supervisor privileges (R2L), probing. The test dataset includes some specific attacks that do not appear in the training dataset to make the task more difficult and realistic, which contains 24 training attack types, with additional 14 types in the test dataset only. KDD CUP99 is mainly used for binary classification (normal and attack) and multiple classification (normal and four kinds of attack).

The following data shows the connection record data format, and each feature records separated by a comma. Each record in the KDD Cup99 data set contains 41 various quantitative and qualitative features which can be divided into three groups: basic features of the network connection, features based on the content of the network connection and features based on time flow in 2 second. The last feature is the label.

2, tcp, smtp, SF, 1684, 363, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0.00, 0.00, 0.00, 0.00, 1.00, 0.00, 0.00, 104, 66, 0.63, 0.03, 0.01, 0.00, 0.00, 0.00, 0.00, 0.00, normal.

0,icmp,ecr_i,SF,1032,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,511,511,0.00,0.00,0.00,0.00,1.00,0.00,0.00,255,255,1.00,0.00,1.00,0.00,0.00,0.00,0.00,0.00,smurf.

0,udp,private,SF,28,0,0,3,0,0,0,0,0,0,0,0,0,0,0,0,0,2,2,0.00,0.00,0.00,0.00,1.00,0.00,0.00,255,2,0.01,0.02,0.01,0.00,0.00,0.00,0.77,0.00,teardrop.

The data has been preprocessed before using for training and testing of the classification model. The preprocessing of dataset has been explained in section 4.B.

B. Data preprocessing

The standard KDD Cup99 dataset is in text format. Some of the 41 features are irrelevant, and some others may be redundant, which can reduce efficiency and lead to wrong results. In this paper, after use of general feature selection techniques for feature simplification, features with the same value and less value are deleted. Finally, features that can improve the classification accuracy and running efficiency of the algorithm are selected. With this, data size reduction by reducing a number of features from 41 to 35 is shown in Table 1.

TABLE I. Features selection

Feature groups	Features
Basic	duration,protocol_type,service,flag,src_bytes,dst_bytes,wrong_fragment
Content based	hot,num_failed_logins,logged_in,num_compromised,root_shell,num_root,num_file_creations,num_access_files,is_guest_login
Time based	count,srv_count,serror_rate,srv_serror_rate,rerror_rate,srv_rerror_rate,same_srv_rate,diff_srv_rate,srv_diff_host_rate,dst_host_count,dst_host_srv_count,dst_host_same_srv_rate,dst_host_diff_srv_rate,dst_host_same_src_port_rate,dst_host_srv_diff_host_rate,dst_host_serror_rate,dst_host_srv_serror_rate,dst_host_rerror_rate,dst_host_srv_rerror_rate

As the dimension of the dataset is quite different which makes the running time longer, it needs to be standardized. The most common standardization method is Z-score which is called zero-mean normalization. After preprocessing, the data conform to the standard normal distribution that the mean is 0 and standard deviation is 1. This is given by:

$$z(x) = \frac{x - \bar{x}}{s(x)} = \frac{x - \bar{x}}{\sqrt{\frac{(x - \bar{x})^2}{n}}} \quad (9)$$

Where x is the original data, $\bar{x}$ is the mean of all data, and n is the number.

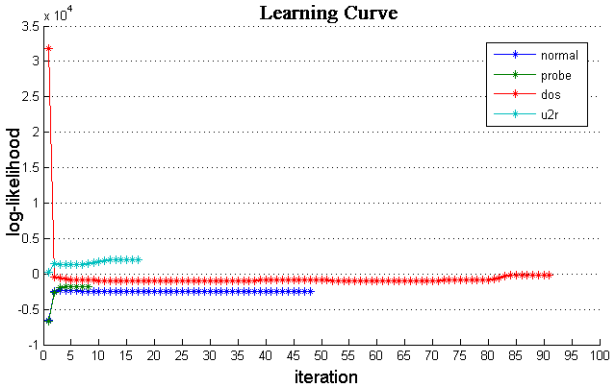
C. Results in preliminary identification layer

After feature selection, 35 features are used to form attack feature set S, which represents the type of attack. According to the classification of features, 35 features can be divided into three groups: Basic, content based, time-based features. According to the four attack types and normal of KDD Cup99, the sample set of each feature subspace is formed. Table 2 shows the number of records for each attack type in the training and test datasets, respectively. The network attack sample set we used is established as R.

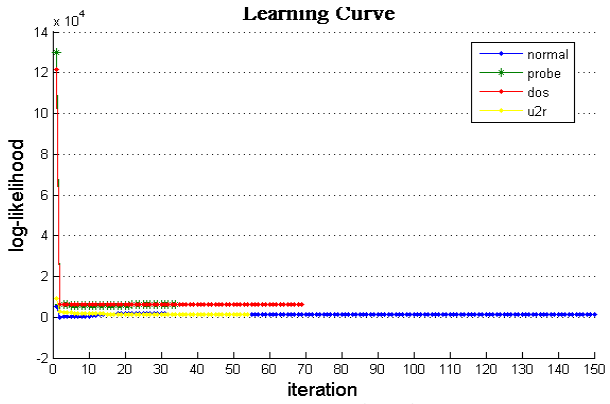
TABLE II. CONNECTION RECORDS OF TRAINING AND TEST SETS

Type of connection	Available training set	Training set	Available test set	Test set
Normal	972780	8000	60593	605
Probe	41102	4000	4166	1359
DoS	3883370	8000	229853	2230
R2L	1126	1126	16189	1618
U2R	52	52	228	228

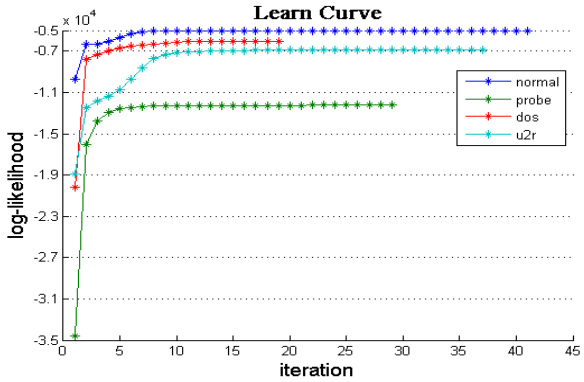
The network connection feature parameter space S is divided into three sub spaces S_j ($j=1, 2, 3$), S_1 is basic features, S_2 is content based features and S_3 is time based features. According to the definition of the feature subspace, feature parameter values are chosen from attack sample set R to consist of attack training sample set of feature subspace FR_{ij}. For each feature subspace S_j, the HMM has been trained for learning. While training the model, it is necessary to initialize appropriate values $\lambda_0 = \{\pi_0, a_0, b_0\}$, as the performance of the model mainly depends on these values. In this paper, initial parameters are generated randomly. After initialization of parameters λ_0, the model selection is a major issue. Standard Baum-Welch algorithm and EM algorithm are used to train the model. The forward algorithm is suitable to test the network traffic. Then the model parameters are $\lambda_1 = \{\pi_1, a_1, b_1\}$ after training. The learning curve of model training is shown in Figure 3.



(a) HMM training process of sub feature space S_1



(b) HMM training process of sub feature space S_2



(c) HMM training process of sub feature space S_3

Fig. 3. The learning curve of model training of sub feature space

After training the parameters, the model has the ability to learn, and is tested by test sample data. A test data of DoS attack is chosen to be evaluated with four kinds of sub-HMMs respectively. The results are shown in Table 3.

TABLE III. LOG LIKELIHOOD VALUE IN PRELIMINARY IDENTIFICATION

	S_1	S_2	S_3	loglik
Normal	-54.57	0.12	-32.4	-86.85
Probe	-33.79	5.31	-11.13	-39.61
DoS	-6.22	5.55	-25.82	-26.49
R2L	-62	-1.3	-31.2	-94.5
U2R	-16.12	-0.97	-55.53	-72.62

Log likelihood(loglik) represents the match value between the test data and the HMM (three parameters: prior1, transmat1, obsmat1). The bigger loglik value means matching better. In table 2, the maximum of the sum of the log-likelihood probability of all sub features is DoS. So, the initial judgment for the test sample is DoS.

D. Results in fusion decision layer

The frame of attack discernment is established as $\Theta = \{A_1, A_2, A_3, A_4, A_5\}$, where A_i is attack type : Normal, Probe, Dos, U2R or R2L.The forward algorithm of evaluation function of HMM is used to calculate the probability of an observed sequence with the given hidden Markov model. The output of each feature model H_{ij} is as the body of evidence. The equation (7) is used to obtain basic probability assign of the proposition

A_i distributed by all evidences. Table 4 shows the performance comparison of this method with classical DS, Yager method, Sun Quan method, and Murphy method. Calculate the average results of 10 times using DoS test dataset. Classification performance of several fusion methods is showed as follows.

TABLE IV. PERFORMANCE COMPARISON OF THE PROPOSED APPROACH WITH OTHER FUSION METHODS

Threshold	Fusion method	Time/s	Classification accuracy%
$\varepsilon_1 = 0.8$ $\varepsilon_2 = 0.1$	DS	4.978446	60.50
	Yager	4.688622	36.30
	Sun Quan	4.872527	79.30
	Murphy	4.748692	83.43
	Our method	4.617725	88.42
$\varepsilon_1 = 0.7$ $\varepsilon_2 = 0.15$	DS	4.662461	72.30
	Yager	4.696050	45.25
	Sun Quan	4.788757	83.01
	Murphy	4.479943	88.40
	Our method	4.607600	93.36
$\varepsilon_1 = 0.6$ $\varepsilon_2 = 0.2$	DS	4.872562	79.00
	Yager	4.953292	53.75
	Sun Quan	4.919677	87.39
	Murphy	4.744361	86.20
	Our method	4.870531	95.83

From the table, it can be concluded that this evidence fusion method has the best classification accuracy with different thresholds compared with other methods. The time required for classification is almost the same. When ε_1=0.8 and ε_2=0.1, the classification accuracy can reach 88.42%, while ε_1=0.6 and ε_2=0.2, the accuracy can reach 95.83%. In addition, in order to compare with the classification results of pure HMM method, traditional hidden Markov model method is applied to classify attacks. Here, the model is trained with training samples from training data set as in the above case. Results of training process is in Figure 4, and Table 5 shows the comparison between HMM-DS and HMM.

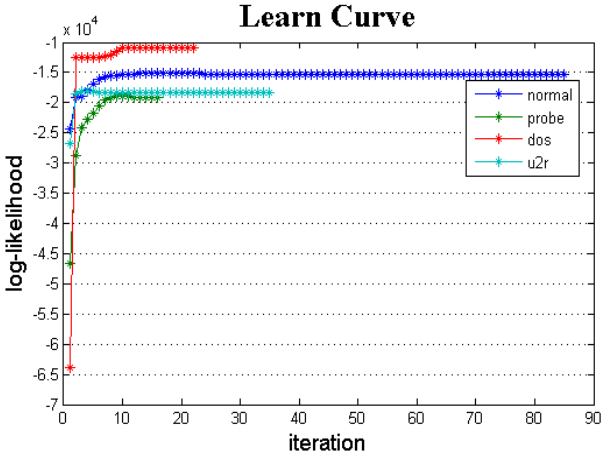


Fig. 4. HMM training process of feature space S

TABLE V. COMPARISON OF LOG LIKELIHOOD AND TIME IN DOS MODEL BUILDING

	Hmm_1	Hmm_2	Hmm_3	Hmm
loglik	-2460.63	-5307.02	-6342.99	-10850.15
Running time/s	20.7	101.56	78.2	273.88

Compare Figure 3 and Figure 4, log likelihood of feature sub space after training is less than the value of the whole feature space. Therefore, parameters are better optimized, and the total training time of HMM-DS is lower than of pure HMM with improved accuracy. Select test samples from each attack test data set. The classification performance of each attack type is shown in Table 6 and Table 7.

TABLE VI. RESULTS OF HMM AND DS

Attack type	Normal	Probe	DoS	R2L	U2R	Classification accuracy (%)
Normal	580	14	6	4	1	95.8
Probe	61	1280	17	1	0	94.1
DoS	122	58	2050	0	0	91.9
R2L	102	1	0	1490	25	89.6
U2R	19	2	0	4	195	88.6

TABLE VII. RESULTS OF PURE HMM

Attack type	Normal	Probe	DoS	R2L	U2R	Classification accuracy %
Normal	564	19	8	10	4	93.22
Probe	132	1359	7	0	0	89.80
DoS	128	100	2002	0	0	89.70
R2L	11	0	0	1125	2	80.90
U2R	91	2	0	11	124	54.40

Compared with pure HMM, HMM-DS system proposed in this paper can significantly improve the classification accuracy and speed. Other evidence fusion methods can improve the speed, but the classification accuracy is low. The comparison results of several common classification methods with the proposed approach are shown in Table 8.

TABLE VIII. COMPARISON BETWEEN THE PROPOSED APPROACH AND COMMON METHODS

Method	Normal	Probe	DoS	R2L	U2R
C4.5	97.08	87.62	96.08	8.12	23.69
LibSVM	91.83	85.26	97.30	18.29	25.88
NB	96.63	89.94	90.20	8.12	24.12
Our method	96.20	95.60	92.30	87.70	89.20

As shown in Table 8, all methods have high classification accuracy of Probe and DoS attack, but common methods are lower of R2L and U2R attack. For some business, and government networks, U2R and R2L attack have more damage than Probe and DoS attack. Thus, higher detection rate of U2R and R2L is equally important with the whole detection rate. Moreover, from the above discussion, it can be noticed the superiority of the proposed HMM-DS over other methods.

In conclusion, reasons for the above results are:

1) The network attack feature parameters space *S* is divided into several sub spaces which can reduce the dimension of the input vector for hidden Markov model. The training speed of each sub-model is accelerated, thus the classification speed of the HMM-DS method is improved.

2) The output results of each sub hidden Markov model are used as bodies of evidence. Some evidences are consistent

while some are conflicting. The evidence method proposed in this paper can effectively fuse these evidences.

3) The input of traditional HMM and other classification methods is ultra-high dimensional feature space. As some features interfere with each other, the speed and accuracy of classification is very low.

V. Conclusion

In this paper, the original feature parameters space of attacks were divided into several sub-feature spaces and a corresponding sub hidden Markov model for each sub-feature space was built. DS evidence theory method was applied to fuse the output of sub hidden Markov model, which can classify attacks effectively. The results show that this fusion system based on HMM-DS is obviously superior to the pure HMM or DS method, and combined the advantages both of HMM and DS. Hence, the proposed approach take advantage of HMM dealing with continuous dynamic signal, and calculate the match value between HMM model and unclassified data to form basic probability assignment which is provided for DS fusion decision. The advantage of DS can make up the shortage of HMM in making maximum probability judgment. the proposed approach proved to work well in combination of all kinds of evidence and to outperform other techniques in terms of classification accuracy. Experiments results show that the proposed approach can improve accuracy and speed of classification.

Although the proposed HMM-DS classification approach looks promising, there is still a large room to improve the classification accuracy for unknown attacks. In order to apply this scheme to other types of classification and recognition problems, a general framework for this approach needed to be constructed.

Acknowledgment

This work was supported by National Natural Science Foundation of China (No. 61202439), and the Fundamental Research Funds for the Central Universities of China.

References

[1] Panda, M., Abraham, A., Das, S., and Patra, M. R.,"Network intrusion detection system: A machine learning approach, " Intelligent Decision Technologies, vol. 5, no.4, pp. 347–356, 1955.

[2] BM Aslahi-Shahri, R Rahmani, M Chizari, A Maralani, M Eslami, MJ Golkar, A Ebrahimi, "A hybrid method consisting of GA and SVM for intrusion detection system" , Neural Computing & Applications, pp. 1-8, 1955.

[3] Cheng Xiang, Png Chin Yong, Lim Swee Meng, "Design of multiple-level hybrid classifier for intrusion detection system using Bayesian clustering and decision trees", Pattern Recognition Letters, vol. 29, no. 7, pp. 918–924 ,2008

[4] Levent Koc, Thomas A. Mazzuchi, Shahram Sarkani, "A network intrusion detection system based on a Hidden Naïve Bayes multiclass classifier", Expert Systems with Applications, vol. 39, no. 18, pp. 13492-13500 ,2012

[5] Yuk Ying Chunga, Noorhaniza Wahid, "A hybrid network intrusion detection system using simplified swarm optimization (SSO), Applied Soft Computing, vol. 12, no. 9, pp. 3014-3022 ,2012

[6] SJ Horng,MY Su,YH Chen,TW Kao,RJ Chen et al., "A novel intrusion detection system based on hierarchical clustering and support vector machines", Expert Systems with Applications, vol. 38, no. 1,pp. 306-313, 2011.

[7] Y Li, J Xia, S Zhang, J Yan, X Ai, "An efficient intrusion detection system based on support vector machines and gradually feature removal method", Expert Systems with Applications, vol. 39,no. 1,pp. 424-430,2012.

[8] D Ariu, R Tronci, G Giacinto, "HMMPayl: An intrusion detection system based on Hidden Markov Models", Computers & Security, vol. 30, no. 4, pp. 221‑241, 2011.

[9] S Jha, K Tan, RA Maxion, "Markov chains, classifiers, and intrusion detection", Computer Security Foundations Workshop, the 14th IEEE, Cape Breton, Novia Scotia, 2001, pp. 206-219.

[10] DEMPSTER A P, "Upper and low probabilities induced by a multi-valued mapping", Annuals of Mathematical Statistics, vol. 38, no. 6, pp. 325-339, 1967.

[11] SHAFER G A, "Mathematical theory of evidence," Princeton: Princeton University Press, 1976.

[12] FRADA B, CLYDE W H, " Handbook on decision support systems ," Heidelberg, Berlin,Springer, 2008.

[13] Chao Ning, Maoyin Chen, and Donghua Zhou, "Hidden Markov Model-Based Statistics Pattern Analysis for Multimode Process Monitoring: An Index-Switching Scheme", Industrial &Engineering Chemistry Research, vol. 53, no. 27, pp. 11084-11095,2014.

[14] W.Z. Wu, "Attribute reduction based on evidence theory in incomplete decision systems", Information Sciences, vol. 178, no. 2008, pp. 1355–1371.

[15] J Weisberg, "Dempster-Shafer Theory", International Journal of Approximate Reasoning, 2010.

[16] Yager R R, "On the Dempster-Shafer framework and new combination rule", Information Sciences, vol. 41, no. 2, pp.93-138, 1987.

[17] Yager RR, "On the fusion of non-independent belief structures", International Journal of General Systems, vol. 38, no. 5, pp. 505-531, 2009.

[18] SUN Quan, YE Xiuqing, GU Weikang, "A new combination rules of evidence theory", Acta Electronica Sinica, vol. 28, no. 8, pp.117-119, 2000.

[19] Thierry Denoeux. Conjunctive and disjunctive combination of belief functions induced by nondistinct bodies of evidence", Artificial Intelligence, vol. 172,no. 2–3 , pp. 234–264, 2008.

[20] Destercke S, Dubois D, " Idempotent conjunctive combination of belief functions: extending the minimum rule of possibility theory", Information Science, vol. 181, no. 18, pp. 3925-3945, 2011.

[21] Haenni R, "Are alternatives to Dempster's rule of combination real alternatives: comments on "about the belief function combination and the conflict management problem", Information Fusion, vol. 3, no. 4, pp. 237-239, 2002.

[22] Murphy C. Combining belief functions when evidence conflicts [J].Decision Support Systems, 2000, 29(1):1-9.

[23] Deqiang Han, Yong Deng, Chongzhao Han, "Sequential weighted combination for unreliable evidence based on evidence variance",Decision Support Systems, ,vol. 56, no. 6, pp. 387-393, 2013.

[24] Peida Xu, Yong Denga, Xiaoyan Sua, Sankaran Mahadevanc, "A new method to determine basic probability assignment from training data ",Knowledge-Based Systems, ,vol. 46, pp. 69-80, 2013.

[25] Xiaoyan Su, Sankaran Mahadevan, Peida Xu, Y Deng, "Handling of Dependence in Dempster‑Shafer Theory", International Journal of Intelligent Systems, ,vol. 30, no. 4, pp. 441‑467, 2015.

[26] HP Corporation, "Objective Attributes Weights Determining Based on Shannon Information Entropy in Hesitant Fuzzy Multiple Attribute Decision Making", Mathematical Problems in Engineering,vol. 2014, no. 1, pp. 1-7,2014.

[27] http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html