Sensors & Transducers, Vol. 159, Issue 11, November 2013, pp. 230-235

Sensors & Transducers

http://www.sensorsportal.com

A Kind of Network Intrusion Detection Algorithm Based on Quantum-behaved Particle Swarm Optimization

[1] Qiang Song, [2] Lingxia Liu
[1] Anyang Institute of Technology, Anyang, 455000, China
[2] Anyang Normal University , Anyang, 455000, China
[1] E-mail: aysq168@163.com

Received: 24 August 2013 /Accepted: 25 October 2013 /Published: 30 November 2013

Abstract: In order to overcomes the drawbacks of fuzzy clustering methods which are sensitive to the initial values and easily trapped into local minima in intrusion detection algorithm, a hybrid algorithm is proposed based on quantum-behaved particle swarm optimization and semi-supervised kernel fuzzy clustering algorithm. This algorithm can supervise and clustering a few labeled data to generate correct model, use this model to guide lots of unlabeled data to clustering, and enlarge the labeled data set. Those data still cannot be labeled, which are clustered by the kernel fuzzy methods based on quantum-behaved particle swarm optimization, and determine mark types. The simulation of KDD CUP 99 data set is implemented to evaluate the proposed algorithm. Comparing to other algorithms, the result shows the proposed algorithm can obtain the ideal error detection rate and false drop rate in the intrusion detection. *Copyright © 2013 IFSA.*

Keywords: Intrusion detection, Quantum-behaved Particle Swarm Optimization (QPSO), Semi-supervised clustering, Kernel function.

1. Introduction

Intrusion Detection (ID) refers to monitoring the running state of network, finding all kinds of attack attempts, aggressive behavior or attack result, to ensure the availability, integrity, and confidentiality of system resources. Machine learning methods being applied into intrusion detection system can make the system have a stronger adaptability, self-study habit and robustness, and it is currently an important direction of intrusion detection research.

Machine learning is divided into supervised learning, unsupervised learning [2] and semi-supervised learning [3]. A semi-supervised learning is divided into semi-supervised classification, semi-supervised clustering, and semi-supervised regression. A semi-supervised clustering [2] is a new clustering method in recent years, which integrates the advantages of no-supervised learning [1] and supervised learning [3], improves the quality of the clustering, is one of the important research direction in the field of machine learning and pattern recognition in recent years. Because the anomaly detection algorithm based on supervised learning requires to access a lot of category information of tag data, and tagged data is relatively limited, a lot of cost need to pay for they; simultaneously anomaly detection algorithm based on unsupervised learning groups according to the similarity of data to overcome the shortcomings of tag samples in supervised learning method, but its accuracy is significantly lower than the supervised detection method. However the advantages of a semi-supervised clustering are in the real application, it is possible to get a small amount of the marked sample data, which can use a small amount of supervised

sample information to guide the samples without label to do clustering. The detection accuracy of current semi-supervised detection algorithm can still be improved, especially for the detection of new attack types.

Aiming to the above problems of anomaly intrusion detection based on machine learning, this paper proposes a semi-supervised kernel fuzzy C-means clustering algorithm based on quantum-behaved particle swarm optimization [4]. Kernel Fuzzy C-means algorithm [5-6] (KFCM) maps the samples to feature space, where uses the FCM algorithm to do cluster analysis. To some extent, overcoming the sensitivity to noise data and outliers can correctly clustering to data distributed in different shapes, overcoming the dependence on the internal shape distribution of data can enhance the algorithm robustness.

But KFCM is as same as K-means and FCM, the clustering performance has the shortcomings of depending on the selection of initial cluster center and easily fall into local optimum, therefore in this paper the quantum-behaved particle swarm optimization is applied to a semi-supervised kernel fuzzy clustering method. At first a few tagged data are supervised and clustered to obtain the correct models, and then these models are used to guide a lot of unlabeled data clustering, expand the tagged data set, finally the data which are still not tagged use the kernel fuzzy C-means algorithm based on quantum-behaved particle swarm nuclear optimization to clustering and determine the tag type. The simulation of KDDCUP 99 experimental data verifies the feasibility and effectiveness of QPSO-SKFCM algorithm.

2. Kernel Fuzzy C-means Clustering Algorithm

Traditional fuzzy C-means clustering algorithm [7-8] is a simple algorithm with fast convergence rate, has been widely used in practice. However, fuzzy C-means clustering algorithm is only suitable for the situation that sample distribution appears Gaussian or group distribution. When sample linearity is inseparable or sample is not Gaussian and none-elliptic distribution, the practical applicability of algorithm is poor [9]. Aiming to this problem, Girolami first put forward the idea of combing the nuclear method and the clustering method [9], the data in pattern space was nonlinearly mapped to the feature space, which increased the linear separable probability of model; that is, expanding the difference among model classes, and in the high dimensional feature space to realize the linear gathering. The basic idea of kernel fuzzy C-means clustering algorithm [6, 9] is: at first, a nonlinear mapping ϕ is used, the samples in the original space are mapping to the kernel space, then in the high dimensional kernel space traditional fuzzy C-means clustering is used. In nuclear space the distance $D_h(\phi(x), m_k^{\phi})$ between any sample $\phi(x)$ and the mean m_k^{ϕ} is as follows:

If $X = \{x_k, k = 1, 2, ..., N\} \in R^e$ is the sample set of waiting for classification, thereinto, N is the sample number; $x_k = \{x_{k1}, x_{k2}, ..., x_{ke}\}$ is the e-dimensional characteristic vectors. $K(2 \le K \le N-1)$ is the supposed clustering number, $C = (c_1, c_2, ..., c_k)$ is K cluster centers, $u_{ij} = (i = 1, 2, ..., K; j = 1, 2, ..., N)$ is the membership functions of the j-th sample to the i-th class, and fulfills the conditions $0 \le u_{ij} \le 1$, $\sum_{i=1}^{k} u_{ij} = 1$. The fuzzy clustering problem is according to the clustering principle functions to solve the K cluster centers C of samples set and membership functions matrix $U = [u_{ij}]_{k*N}$. The clustering principle functions of FCM can be represented as:

$$J(U, C) = \sum_{i=1}^{K} \sum_{j=1}^{N} u_{ij}^{m} \| x_j - c_i \|^2 \tag{1}$$

In the expression, $m > 1$ is the weighted index. The purpose of clustering is to make $J(U, C)$ minimum.

The non-linear mapping is defined as $\phi : X \to F$, then $x \in R^e \to \phi(x) \in R^q$, thereinto, F is the high-dimensional characteristic space. The clustering principle function of KDCM is expressed as:

$$\sum_{i=1}^{K} \sum_{j=1}^{N} u_{ij}^{m} \| \phi(x_j) - \phi(c_i) \|^2 = \sum_{i=1}^{K} \sum_{j=1}^{N} u_{ij}^{m} d_F^2(x_j, c_i) \tag{2}$$

$$d_F(x_j, c_i) = \| \phi(x_j) - \phi(c_i) \| = \sqrt{K(x_j, x_j) - 2K(x_j, c_i) + K(c_i, c_i)} \tag{3}$$

In the expression, $K(x_j, C_i)$ is the kernel function; $d_F(x_j, c_i)$ is the Euclidean distance in characteristic space. When kernel function adopts Gaussian score, the clustering principle is simplified as:

$$J(U, C) = \sum_{i=1}^{K} \sum_{j=1}^{n} u_{ij}^{m} (2 - 2K(x_j, c_i)) \tag{4}$$

In the expression,

$$K(x_j, c_i) = \exp[\frac{-\| x_j - c_i \|^2}{\delta^2}], \tag{5}$$

δ is the Gaussian parameter.

The Lagrange multiplier method can solve the expression (4):

$$u_{ij} = \frac{[1-K(x_j,c_i)]^{1/(m-1)}}{[\sum_{i=1}^{k}[1-K(x_j,c_i)^{-1(m-1)}]]^{-1}} \quad (6)$$

$$c_i = \frac{\sum_{j=1}^{n} u_{ij}^m K(x_j,c_i)x_j}{\sum_{j=1}^{n} u_{ij}^m K(x_j,c_i)} \quad (7)$$

3. Quantum-behaved Particle Swarm Optimization Algorithm

3.1. The Particle Swarm Optimization (PSO) Algorithm

Particle Swarm Optimization algorithm is put forward by Dr. J. Kennedy and Dr. Eberhan [10], it is based on the evolution search technology of the population, but all the basic and the improved PSO algorithm can't guarantee the global convergence of the algorithm, the basic particle swarm optimization algorithm model shows the particle flying speed is equal to the search step length, the size directly affects the global convergence of the algorithm.

3.2. The Quantum-behaved Particle Swarm Optimization (QPSO) Algorithm

In 2004, aiming to the convergence problem of PSO algorithm, Sun *et al.* put forward a new model of PSO algorithm [11] from the perspective of quantum science. This model is based on DELTA trap, the particles are regarded as having quantum's actions, according to this model the quantum-behaved particle swarm optimization (QPSO) is proposed, and the experimental result proves that QPSO convergence has greatly improved [11]. In QPSO, the main particles iterative formulas are as follows:

$$X(t+1) = P + \beta * |mbest - X(t)| * \ln(1/u), u = rand \quad (8)$$

$$mbest = \sum_{i=1}^{M} P_i / M = (\sum_{i=1}^{M} P_{i1}/M, \sum_{i=1}^{M} P_{i2}/M, ..., \sum_{i=1}^{M} P_{iD}/M) \quad (9)$$

$$p_{id} = \phi \times P_{id} + (1-\phi) \times P_{gd}, \phi = rand \quad (10)$$

$$\beta = (\beta_1 - \beta_2) \times (MAXITER - t)/MAXITER + \beta_2 \quad (11)$$

In the expressions, β is the coefficient of contraction and expansion, it is an important parameter of QPSO convergence, β_1 and β_2 are initial value and final value of β respectively. P_i is *mbest* of the *i*-th particle, P_g is the global optimal, M is the particles number, D is the particles dimension, and MAXITER is the maximum iteration number, t is current iteration number. From literature [4], when β is greater than 1.7 , it can not make QPSO algorithm convergence, in general, β_1 and β_2 taking 1.2 and 0.7 respectively can make the algorithm achieve better convergence.

4. A Semi-supervised Fuzzy C-means Clustering Algorithm Based on Quantum Particle Swarm Optimization

4.1. The Particle Coding

In particle cluster algorithm, each particle is composed of K cluster centers, the dimensions of the sample vector is S, therefore particles are represented as $S \times K$ dimensional vectors, particle position X_p is constructed as follows:

$$X_p = (c_{p1}, c_{p2}, ..., c_{pi}, ..., c_{pk})$$

In the expression, c_{pi} is the i-th cluster center of the p-th particle.

4.2. Design of Fitness Function

Particle fitness function *f(x)* is defined as the target function *J(U,C)* of KFCM algorithm, that is:

$$f(x_p) = J(U,C) = \sum_{i=1}^{K}\sum_{j=1}^{n} u_{ij}^m (2 - 2K(x_j,c_i)) \quad (12)$$

4.3. The Flowchart of Semi-supervised Kernel Fuzzy Cluster Algorithm Based on Quantum-behaved Particle Swarm Optimization

Algorithm: quantum-behaved particle swarm optimization semi-supervised kernel fuzzy C-means algorithm (QPSO-SKFCM)

Input: the labeled data set $S_{label} = \{(x_i, l_i) | i = 1, 2, ...n\}$, the unlabeled data set $S_{unlabel} = \{x_i | i = 1, 2, ...m\}$, $n << m$, data set $S = S_{label} \bigcup S_{unlabel}$.

Output: the data type of data $x \in S_{unlabel}$ (intrusion or normal)

1) *To supervise and clustering the* S_{label} *, initialize each cluster center* O_k*, the maximum radius of each cluster is* R_k*.*

2) *Initially clustering the samples like* $x \in S_{unlabel}$, *if they can be correctly classified, then they will be put into the cluster* C_k, *otherwise they will be put into* C_r.

for i=1 to m do

calculate the distance between each sample $x \in S_{unlabel}$ *and* O_k, *note as* $r_i, i = 1, 2, ..., k$

end for

$r' = \min(r_i \mid i = 1, 2, ..., k)$

While $r' \le R_k$

Then $x \in C_k$

or $x \in C_r$

end

3) *While the undetected data set* C_r *is empty,*

- *Using the cluster center* O_k, *to produce the initial membership matrix in data set* C_r.
- *Quantum-behaved particle swarm optimization using expression (12) to calculate the fitness of each particle;*

Initialize P_i, P_g;

Using expressions (8), (9), (10), (11) to produce the new generation individuals, and update P_i, P_g.

- *Kernel fuzzy C-means cluster algorithm*

P_g *is regarded as the initial cluster center*

Expression (6) is used to update the membership matrix; expression (7) can obtain the new cluster center P_s.

- *If* $f(P_s) < f(P_g)$,

then P_g *is used to replace* O_k *and acts the new cluster center.*

else P_s *is used to replace* O_k *and acts the new cluster center.*

end if

end

4) According to the final membership matrix to determine the category of each sample (intrusion or normal).

5. Experiments and Analysis

In order to evaluate the application effect in intrusion detection of a semi-supervised kernel fuzzy clustering algorithm with quantum-behaved particle swarm optimization, KDD CUP99 network data set is chosen as test data set in intrusion detection, which is widely used in the field of intrusion detection field. Experimental data contains four main types:

1) DoS, denial of service attack;

2) U2R, unauthorized access to the local super user privileges;

3) Probe, scanning and detection behavior;

4) R2L, unauthorized access to the remote host.

5.1. The Pretreatment of Experimental Data

The data set in the experiment has total 41 feature attributes, in these features there are many redundant features, only the 21 ones are chosen as the research objects which can reflect part of features of user behavior, respectively they are: Duration, Service (network service at destination terminal), Protocol type, Flag (label of connection correct or wrong), etc.

In order to fulfill the needs of two assumptions in the detection algorithm, four groups of data are selected as the experimental data from the whole detection data set, each group data contains 110029 pieces of records, thereinto, 108929 pieces are normal data, 1100 pieces are intrusion data, and normal data accounts for 99 % in each group of the experimental data, which satisfy the hypothesis of normal data should be far more than the invasion data in detection algorithm.

5.2. Test Result and Analysis

The performance of intrusion detection algorithm mainly considers detection rate and false positive rate (FPR). Through these two indexes the detection result of algorithm can be effectively measured.

$$DR = \frac{Detected\ intrusion\ test\ samples}{Sum\ of\ intrusion\ samples}$$

$$FPR = \frac{the\ normal\ samples\ being\ misreported\ as\ the\ invasion}{the\ normal\ samples}$$

In the experiment, all the relevant parameters are shown in Table 1, the experiment result is shown in Table 2.

From Table 2, the average detection rate of samples reaches 87.13 % using QPSO-SKFCM algorithm, the rate of false positives is only 0.89 %. This suggests it is feasible that a semi-supervised kernel fuzzy clustering algorithm based on quantum-behaved particle swarm optimization is applied to the intrusion detection.

The performances of detecting the intrusion in the data set are compared among QPS-SKFCM algorithm, PSO-FCM algorithm, and FCM algorithm. The experimental results are shown in Table 3.

Table 3 shows in QPSO-SKFCM algorithm the average detection rate reaches 89.7 % to known intrusion, and the average detection rate reaches 81.9 % for unknown intrusion, comparing with other two algorithms, which can better detect intrusion behavior. In the unknown attack detection, the detection to R2L by this algorithm is lower than other type of intrusion detection; this is because R2L invasion is disguised as a legal user's identity to attack, making its characteristics similar with the normal packet, which is easy to cause detection more difficult.

Compared with other two algorithms, in general the algorithm has obvious advantages in detection of unknown and known intrusion type.

Table 1. The selection of experimental parameters.

Parameter Name	Value	Selection Reasons
MAXITER	250	The result has no changes after the iteration number is greater than 250
β_1	1.2	According to literature[11]
β_2	0.7	According to literature[11]
u, ϕ	[0, 1]	Using *rand* functions to pick up one random in [0, 1]
M	50	The selected particle number according to the experiment
K	70	The selected cluster number according to the samples
σ	150	The default value

Table 2. The test results to intrusion by QPSO-SKFCM algorithm (%).

Test data	Detection rate	Misreporting rate
Test data in the first group	87.63	0.91
Test data in the second group	84.20	1.24
Test data in the third group	90.28	0.69
Test data in the fourth group	86.41	0.73
Average	87.13	0.89

Table 3. Comparisons between QPSO-SKFCM algorithm and other algorithms.

Detection algorithm	Known intrusion				Unknown intrusion			
	Dos	U2R	Probe	R2L	Dos	U2R	Probe	R2L
QPSO-SKFCM	82.3	87.4	98.7	87.5	78.9	82.1	93.5	73.2
PSO-FCM	50.2	83.6	96.1	90.2	62.4	74.3	85.7	67.4
FCM	32.9	72.7	86.5	93.6	11.2	47.8	21.5	74.6

6. Conclusions

Experimental results show that detection rate of QPSO-SKFCM algorithm to known intrusion is not only close to intrusion detection method based on supervision, but also the detection rate to unknown attack is higher than the intrusion detection method based on non-supervision. Because QPSO-SKFCM algorithm using a few labeled samples to produce correct sample model to guide lot of unlabeled samples to clustering with supervision, for those unlabeled samples after clustering, kernel fuzzy C-means algorithm via quantum-behaved particle swarm optimization proceeds the unsupervised clustering to realize the detection of unknown attacks, and effectively compensate for the shortcomings of intrusion detection algorithm purely based on supervised learning or unsupervised learning. Table 3 shows the detection of FCM to R2L is still slightly higher than the algorithm in this paper, and the algorithm in this paper at first will need to specify the number of particles and the parameters β_1 and β_2 in advance, selecting the appropriate value has very important influence on algorithm's speed and detection accuracy, this is also the direction for further research.

References

[1]. Weston J., Watkins C., Support vector machines for multi-class pattern recognition, Department of Computer Science, Royal Holloway, *University of London,* 1998.

[2]. Basu S., Baneriee A., Mooney R., Semi-supervised clustering by seeding, in *Proceedings of the 19th International Conference on Machine Learning*, *Morgan Kaufmann Publishers*, San Francisco, CA, 2002, pp. 19-26.

[3]. Flanagan J. A., Unsupervised clustering of symbol strings, *Inter National Joint Conference on Neural Networks,* Portland, Oregon, USA, 2003, pp. 3250-3255.

[4]. Guangquan Yang, Changming Zhu, Kernel fuzzy clustering algorithm based on particle swarm optimization, *Journal of Shanghai Jiaotong University*, 43, 6, 2012, pp. 101-104.

[5]. Kaijing Chan, Huaitie Xiao, The nuclear C-means clustering algorithm selected by initial clustering center optimization, *Computer Simulation*, 26, 7, 2009, pp. 118-121.

[6]. Qingfeng Pan, Shuili Chen, Guolong Chen, Fuzzy C-means clustering algorithm based on kernel functions, *Journal of Jimei University*, 11, 4, 2009, pp. 21-24.

[7]. Dapeng Sun, The research of improved fuzzy clustering algorithm in intrusion detection, *Journal of Computer and Digital Engineering*, 20, 3, 2009, pp. 88-91.

[8]. Lingjie Zhang, Guohui Zhang, Intrusion detection research based on hybrid particle swarm optimization, *Journal of Computer Applications and Software*, 26, 4, 2009, pp. 21-25.
[9]. Zhaoqi Bian, Xuegong Zhang, Pattern Recognition, 2 edition, *Tsinghua University Press,* Beijing, 2000.
[10]. Kennedy J., Eberhart R. C., Particle swarm optimization, in *Proceedings of the IEEE International Joint Conference on Neural Networks*, 1995, pp. 1942-1948.
[11]. Sun J., Xu W. B., A global search strategy of quantum-behaved particle swarm optimization, in *Proceedings of IEEE Conference on Cybernetics and Intelligent Systems*, 2004, pp. 111-116.
[12]. Jihong Pei, Jiulun Fan, Weixin Xie, A new efficient soft clustering algorithm, *Journal of Electronics*, 26, 2, 2008, pp. 83-86.