Hindawi Publishing Corporation
Journal of Sensors
Volume 2016, Article ID 6047023, 11 pages
http://dx.doi.org/10.1155/2016/6047023

Research Article

Multitask Learning-Based Security Event Forecast Methods for Wireless Sensor Networks

Hui He,[1] Dongyan Zhang,[2] Xing Wang,[1] Min Liu,[3] Weizhe Zhang,[1] and Junxi Guo[4]

[1]*School of Computer Science and Technology, Harbin Institute of Technology, Harbin, Heilongjiang 150001, China*
[2]*Department of Software Engineering, University of Science and Technology Beijing, Beijing 100083, China*
[3]*Network Information Center, Harbin Institute of Technology, Harbin, Heilongjiang 150001, China*
[4]*School of Software, Harbin Institute of Technology, Harbin, Heilongjiang 150001, China*

Correspondence should be addressed to Dongyan Zhang; zhangdy@ustb.edu.cn

Received 17 December 2015; Accepted 17 February 2016

Academic Editor: Fei Yu

Copyright © 2016 Hui He et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Wireless sensor networks have strong dynamics and uncertainty, including network topological changes, node disappearance or addition, and facing various threats. First, to strengthen the detection adaptability of wireless sensor networks to various security attacks, a region similarity multitask-based security event forecast method for wireless sensor networks is proposed. This method performs topology partitioning on a large-scale sensor network and calculates the similarity degree among regional subnetworks. The trend of unknown network security events can be predicted through multitask learning of the occurrence and transmission characteristics of known network security events. Second, in case of lacking regional data, the quantitative trend of unknown regional network security events can be calculated. This study introduces a sensor network security event forecast method named Prediction Network Security Incomplete Unmarked Data (PNSIUD) method to forecast missing attack data in the target region according to the known partial data in similar regions. Experimental results indicate that for an unknown security event forecast the forecast accuracy and effects of the similarity forecast algorithm are better than those of single-task learning method. At the same time, the forecast accuracy of the PNSIUD method is better than that of the traditional support vector machine method.

1. Introduction

Sensor network is a network system that integrates monitoring, control, and wireless communication and has a high node number (thousands or even tens of thousands) and dense node distribution. Owing to environmental influences and energy depletion, nodes easily break down. Environment interference and node fault change the network topological structure. Wireless sensor networks have strong dynamics and uncertainty, including changes in network topology, node disappearance or addition, and various threats. Therefore, wireless sensor networks should have strong adaptability to various security attacks so that even if one attack behavior succeeds, the influence of such an attack will only be the minimum. However, attackers can cause sensor networks to be in a partial or total paralysis state through fake and signal interference and thus destroy the system availability, such as through the denial of service attack.

A large-scale sensor network has several advantages. The information from different space perspectives has a high information-price ratio. The distributed processing of a large amount of acquired information can improve the monitoring accuracy and decrease the accuracy requirement of a single-node sensor. Numerous redundant nodes provide the system with strong fault-tolerance performance. Numerous nodes can also increase the covered monitoring area and decrease the cave and dead zones.

Studies on the security situation of wireless sensor networks are presently lacking, but some research results have been obtained on common network security situation evaluation and forecast. The results are derived from *three aspects, namely, model, knowledge representation, and assessment methods*.

First, from the *model aspect*, based on data from the Computer Oracle and Password System, Ortalo et al. [1] adopted privilege graph theory to model system bug and

hole and utilized the Markov model to calculate the mean costs when attackers beat the system security objectives and quantitatively measure system safety. Feng et al. [2] utilized vulnerability scanning and other technological means and quantized the evaluation indexes of vulnerability factors to summarize the standards, methods, tools, and models in the field of information security risk evaluation. Xiao and Dai [3] applied a multilayer fuzzy comprehensive evaluation method to evaluate the information system risk level. Tao [4] and Wang et al. [5] proposed an immunity-based quantitative detection model and method for network security risk. On the basis of this model, Wang et al. [6] introduced a fault model-based risk evaluation model for network information system and the formative description of risk management. Yong et al. [7] proposed a multiangle quantitative model as the evaluation system framework and support platform of a network information system security test. Wu et al. [8] provided an efficiency-evaluating approach for the security measures of an information system under the given vulnerability set. This approach employed colored Petri-Net tools for uniform modeling and simulated the interaction among the workflow, attack flow, and security measures of the system.

Second, from the *knowledge representation aspect*, Yan [9] presented concepts of composition independence, combination complementary, and combination relevance security factors. A formative evaluation model of information system security measurement and its realization were provided through a definition of the correlation among access path, standard path, and components. Some methods combined qualitative and quantitative methods for studies on network security evaluation through evaluating the mutual effects of threat, bug, and attack results. For example, the multistage attack model framework proposed by Clark et al. [10] applied vulnerability description, object-oriented network modeling, and attacker expression capacity to carry out qualitative and quantitative complicated vulnerability risk analyses. Zhao et al. [11] also proposed risk level formula and entropy weight, applied likelihood estimation of risk probability, and utilized entropy theory to handle the weight vector of risk factors specific to the analytic hierarchy process and the strong subjectivity of fuzzy logic.

Finally, from the *assessment methods*, Bilar [12] compared the security event data caused by the modified time of bug and patch and quantitatively analyzed the risk of information system network. Cardoso and Freire [13] proposed a quantification evaluation method of an agent-based distributed system. An agent was utilized to monitor the host security statuses of a subnet or other nodes for evaluation. Chen et al. [14] introduced a quantitative evaluation method of service and host-to-network stratification network security threat situation, which further expanded the bug-based vulnerability analysis to the network level. Wang et al. [15, 16] designed a multisource fusion evaluation system of network security situation awareness. Eom et al. [17] introduced asset indicators of the assessed information system. Machine learning in China and abroad mainly forecasts situations through a neural network method. Shanghai Jiaotong University and Harbin Engineering University, respectively, realized situation forecasts through a radial basis function [18] and a genetic algorithm, back propagation neural network, as well as providing a preliminary solution to the situation forecast problem. Anwar et al. [19] proposed an index system that converged network equipment, service, vulnerability, access control, and other aspects. Kandula et al. [20] considered fine granularity to assess the problems of the enterprise network—host process level. Gong and Zhuo [21] introduced cyberspace situational awareness (CSA) and its origin, conception, objective, and characteristics. This study provided the development directions of CSA and offered conclusions from issue, technical, and application systems. Recently, [22, 23] proposed a network threat assessment based on alert verification. Tian et al. [24] also proposed a method to quantitatively assess network threat situations based on alerts and contextual information.

For the security event attack features of wireless sensor networks [25], a multitask learning method is adopted for forecasting. This method uses domain information in the training signals of relevant tasks for inductive bias and further improves the generalization performance. Multitask learning refers to the parallel learning of learning tasks with sharing expression, and the learning of each task can help improve the learning of other tasks. Accordingly, multitask learning has better effects on machine learning by a single neural network.

On the whole, the study on wireless sensor network security in this paper is of considerable importance. On the one hand, a large-scale sensor network is used for topology partitioning and for calculating the similarity degree among regional subnets. The occurrence and development trends of unknown network security events can be forecasted and speculated through multitask learning of the occurrence and transmission characteristics of the known network security events, thereby providing bases for the effective control of unknown network attacks and restraining the entire network transmission of attacks. On the other hand, given that many subregions in the network have no security detection methods, the attack data in these regions may not be collected. In this case, part data from this region or data from other similar subregions are utilized to forecast network attacks in data-missing regions. The quantitative evaluation value of unknown regional network security events can also be calculated. This study adopts the PNSIUD method in case of data missing. The method employs the basic idea of the transductive support vector machine (TSVM) method to predict the missing attack data in the target region according to the known data in similar regions.

2. Region Similarity-Based Security Event Forecast for Wireless Sensor Networks

When large-scale security events occur in each subregion of wireless sensors, a situation forecast needs to be made in each subregion of the network according to requirements. However, if a situation forecast is made on the data in this subregion in isolation, it may fail to consider the influences of the surrounding network on the target network and the similarities in other subregions. Hence, the forecast results will not be ideal. In this way, the measurement data from other similarity subregions can be utilized to improve the situation evaluation of the target network and the forecast accuracy.

Table 1: Similarity measurement index of the network topological structure.

Network structure (NS)	Topology (Topo)	Nodes
		Edges
		Slope of degree distribution linefit (SDDL)
		Normalized average neighbor degree (NAAND)
		Slope of coreness distribution linefit (SCDL)
		Assortative coefficient (AC)
		Max edge betweenness (MEB)
	Infrastructure (Inf)	Standard deviation of distance (SDD)
		Distribution of bandwidth (DB)
		Distribution of delay (DD)
Network traffic (NT)	Proportion of traffic classification (PTC)	
	Proportion of total traffic (PTT)	
	Traffic distribution within the network (TDN)	
Asset	Number of important assets (NIA)	
	Important asset value (IAV)	
	Position distribution of assets (PDA)	
Security appliances (SA)	Traffic monitor (TM)	Number of TM (NTM)
		Position distribution of TM (PDTM)
		Type (TTM)
	Intrusion detection system (IDS)	Number of IDS (NIDS)
		Position distribution of IDS (PDIDS)
		Type of IDS (TIDS)
	Network management (NM)	Type of NM (TNM)

This study adopts a multitask learning method to forecast the situation of the target network. This method fully utilizes data from the surrounding regions of the target network and data from similar regions in the network environment of the target region.

Based on the aforementioned idea, this study forecasts two targets, namely, the topology element-based forecast on target network availability damage and the overall situation forecast of target network security events.

2.1. Similarity Measurement of the Sensor Network. This study utilizes measurement data from other similar subregions in the network as an auxiliary knowledge structure to improve the evaluation of the target region and the forecast accuracy. All similarities of the network are measured. Network similarity measures the similarity of two network environments. If two network environments are similar, then the knowledge structures of data between the two networks may be similar, and they can transfer knowledge to realize shared knowledge.

The coexistence of a heterogeneous environment, multiple vendors, and multiple protocols presently leads to a highly complicated network environment and the gradual diversification of user demands. In this case, the network similarity measurement will be extremely complicated. First, in selecting network similarity indexes, various influence factors of network similarity exist; thus, selecting the proper measurement indexes of target problems is important. Second, after proper measurement indexes are selected, guaranteeing the efficiency and correctness of the similarity calculation is crucial. To satisfy the requirements of quantitative evaluation and forecast of a large-scale network, this study selects proper network similarity measurement indexes and formalizes and defines the calculation process of network similarity measurement, which establishes a good foundation for forecasting the region similarity-based security event of a national large-scale network.

2.1.1. Measurement Indexes of Network Similarity. Through analysis and argument, we consider that the influence factors of the heterogeneous degree of a large-scale network mainly include network structure, infrastructure, assets, network security device, and network flow information. Each part can be divided into concrete influence factors. Topology vulnerability factor is selected as a measurement index of the overall similarity measurement of the network topological structure. The overall measurement is shown in Table 1.

2.1.2. Design of the Similarity Measurement of Wireless Sensor Subnetworks. The similarity of two networks is formally defined according to the overall measurement indexes. Supposing the two networks are P and Q, $P = (\mathrm{NS}_P, \mathrm{NT}_P, \mathrm{Asset}_P, \mathrm{SA}_P)$ and $Q = (\mathrm{NS}_Q, \mathrm{NT}_Q, \mathrm{Asset}_Q, \mathrm{SA}_Q)$.

First, the vector of the P network is set as follows:

(1) The network structure similarity measurement is $\mathrm{NS}_p = (\mathrm{Topo}_p, \mathrm{Inf}_p)$, where the network topological structure is $\mathrm{Topo}_p = (\mathrm{Node}_p, \mathrm{Edges}_p, \mathrm{SDDL}_p, \mathrm{NAAND}_p, \mathrm{SCDL}_p, \mathrm{AC}_p, \mathrm{MEB}_p, \mathrm{SDD}_p)$, and the network infrastructure is $\mathrm{Inf}_p = (\mathrm{DB}_p, \mathrm{DD}_p)$.

(2) The network flow similarity measurement is $NT_p = (PTC_p, PTT_p, TDN_p)$.

(3) The network asset similarity measurement is $Asset_p = (NIA_p, IAV_p, PDA_p)$.

(4) The network security equipment similarity measurement is $SA_p = (TM_p, IDS_p, NM_p)$, where the flow monitoring is set as $TM_p = (NTM_p, PDTM_p, TTM_p)$, the intrusion detection is $IDS_p = (NIDS_p, PDIDS_p, TIDS_p)$, and the network management is $NM_p = (TNM_p)$.

Second, the vector of the Q network is set as follows:

(1) The network structure similarity measurement is $NS_Q = (TOPO_Q, Inf_Q)$, where the network topological structure is $Topo_Q = (Node_Q, Edges_Q, SDDL_Q, NAAND_Q, SCDL_Q, AC_Q, MEB_Q, SDD_Q)$, and the network infrastructure is $Inf_Q = (DB_Q, DD_Q)$.

(2) The network flow similarity measurement is $NT_Q = (PTC_Q, PTT_Q, TDN_Q)$.

(3) The network asset similarity measurement is $Asset_Q = (NIA_Q, IAV_Q, PDA_Q)$.

(4) The network security equipment similarity measurement is $SA_Q = (TM_Q, IDS_Q, NM_Q)$, where the flow monitoring is set as $TM_Q = (NTM_Q, PDTM_Q, TTM_Q)$, the intrusion detection is $IDS_Q = (NIDS_Q, PDIDS_Q, TIDS_Q)$, and the network management is $NM_Q = (TNM_Q)$.

2.1.3. Formalized Network Similarity Measurement

(1) Overall Network Environment Similarity. Consider the following:

$$S_{\text{Network}}(P,Q) = S_{\text{NS}}(P,Q) + S_{\text{NT}}(P,Q) + S_{\text{Asset}}(P,Q) + S_{\text{SA}}(P,Q). \quad (1)$$

(2) Network Structure Similarity. The network structure similarity is $S_{\text{NS}}(P,Q) = \omega_1 S_{\text{Topo}}(P,Q) + \omega_2 S_{\text{Inf}}(P,Q)$, where $VS_{\text{Topo}} = (S_{\text{Nodes}}(P,Q), S_{\text{Edges}}(P,Q), S_{\text{SDDL}}(P,Q), S_{\text{NAAND}}(P,Q), S_{\text{SCDL}}(P,Q), S_{\text{AC}}(P,Q), S_{\text{MEB}}(P,Q), S_{\text{SDD}}(P,Q))$; and the network topological structure similarity is $S_{\text{Topo}}(P,Q) = \sum VS_{\text{Topo}}\omega_i / \sum \omega_i$, where ω_i is the weight of similarity measurement i. The influence of different node numbers of the network router on the network similarity is not linear. Thus, exponential transformation is performed on the network similarity, and the similarity of node numbers of the network router can be expressed as

$$S_{\text{Nodes}}(P,Q) = \begin{cases} M, & \left|\text{Nodes}_P - \text{Nodes}_Q\right| \le \alpha, \\ L + (M-L)\exp\left(-k\left(\left|\text{Nodes}_P - \text{Nodes}_Q\right| - \alpha\right)\right), & \left|\text{Nodes}_P - \text{Nodes}_Q\right| > \alpha. \end{cases} \quad (2)$$

Compared with the influences of the node number of the router, the influence of connected sides on the network similarity is relatively moderate. Power function transformation is conducted, and the similarity is

$$S_{\text{Edges}}(P,Q) = \begin{cases} M, & \left|\text{Edges}_P - \text{Edges}_Q\right| \le \alpha, \\ L + \dfrac{(M-L)}{\left(1 + k\left(\alpha - \left|\text{Edges}_P - \text{Edges}_Q\right|\right)^{\beta}\right)}, & \left|\text{Edges}_P - \text{Edges}_Q\right| > \alpha, \end{cases} \quad (3)$$

where L is usually set to 0, M is set to 1, and the result is normalized. α represents the node tolerance, and k and β represent the coordinate coefficients used to adjust the similarity decay rate.

Network infrastructure similarity can be concluded as its weighted Euclidean distance:

$$S_{\text{Inf}}(P,Q) = \sqrt{\omega_1 \left|\text{DB}_P - \text{DB}_Q\right|^2 + \omega_2 \left|\text{DD}_P - \text{DD}_Q\right|^2}. \quad (4)$$

(3) Network Flow Similarity. Consider the following:

$$S_{\text{NT}}(P,Q) = \omega_1 S_{\text{PTC}}(P,Q) + \omega_2 S_{\text{PTT}}(P,Q) + \omega_3 S_{\text{TDN}}(P,Q). \quad (5)$$

The similarities of flow classification ratio, exit, and entrance total flow rate and network flow distribution can be measured by covariance:

$$\begin{aligned} S_{\text{PTC}}(P,Q) &= E\left(\left(\text{PTC}_P - \mu_P\right)\left(\text{PTC}_Q - u_Q\right)\right), \\ S_{\text{PTT}}(P,Q) &= E\left(\left(\text{PTT}_P - \mu_P\right)\left(\text{PTT}_Q - u_Q\right)\right), \\ S_{\text{TDN}}(P,Q) &= E\left(\left(\text{TDN}_P - \mu_P\right)\left(\text{TDN}_Q - u_Q\right)\right). \end{aligned} \quad (6)$$

(1) Asset Similarity. Consider the following:

$$\begin{aligned} S_{\text{Asset}}(P,Q) &= \omega_1 E\left(\left(\text{NIA}_P \cdot \text{IAV}_P - \mu_P\right)\left(\text{NIA}_Q \cdot \text{IAV}_Q - u_Q\right)\right) \\ &\quad + \omega_2 E\left(\left(\text{PDA}_P - \mu_P\right)\left(\text{PDA}_Q - u_Q\right)\right). \end{aligned} \tag{7}$$

(2) Network Security Equipment Similarity. Consider the following:

$$\begin{aligned} S_{\text{SA}}(P,Q) &= \omega_1 S_{\text{TM}}(P,Q) + \omega_2 S_{\text{IDS}}(P,Q) \\ &\quad + \omega_3 S_{\text{NM}}(P,Q), \end{aligned} \tag{8}$$

where

$$\begin{aligned} S_{\text{ITM}}(P,Q) &= \sqrt{\frac{\omega_1 S_{\text{NTM}}(P,Q)^2 + \omega_1 S_{\text{PDTM}}(P,Q)^2 + \omega_1 S_{\text{TYPE}}(P,Q)^2}{\sum \omega}}, \\ S_{\text{NTM}}(P,Q) &= \begin{cases} M, & \left|\text{NTM}_P - \text{NTM}_Q\right| \le \alpha, \\ L + (M-L)\exp\left(-k\left(\left|\text{NTM}_P - \text{NTM}_Q\right| - \alpha\right)\right), & \left|\text{NTM}_P - \text{NTM}_Q\right| > \alpha, \end{cases} \\ S_{\text{PDTM}}(P,Q) &= E\left(\left(\text{PDTM}_P - \mu_P\right)\left(\text{PDTM}_Q - u_Q\right)\right), \\ S_{\text{IDS}}(P,Q) &= \sqrt{\frac{\omega_1 S_{\text{NIDS}}(P,Q)^2 + \omega_1 S_{\text{PDIDS}}(P,Q)^2 + \omega_1 S_{\text{TIDS}}(P,Q)^2}{\sum \omega}}. \end{aligned} \tag{9}$$

The influences of location number difference in the network intrusion detection system (IDS) on network similarity are not in linear relation. Thus, exponential transformation is conducted to obtain the network IDS location number similarity. Consider

$$S_{\text{NIDS}}(P,Q) = \begin{cases} M, & \left|\text{NIDS}_P - \text{NIDS}_Q\right| \le \alpha, \\ L + (M-L)\exp\left(-k\left(\left|\text{NIDS}_P - \text{NIDS}_Q\right| - \alpha\right)\right), & \left|\text{NIDS}_P - \text{NIDS}_Q\right| > \alpha, \end{cases} \tag{10}$$

where L is 0, M is 1, and a represents the node tolerance. Given that the IDS number in the actual network is small, the value of α is set small, whereas k is set large to increase the similarity decay speed.

2.2. Region Similarity-Based Security Event Damage Prediction for Wireless Sensor Network. A multitask learning method is used to forecast; it utilizes the domain information in dependent-task training signals for induction bias and improving generalization performance. Multitask learning performs parallel learning on multiple learning tasks with sharing expression, and the learning of each task can help other tasks learn better. Furthermore, it has better effects than single-task machine learning.

People have proposed an algorithm that considers the connections among subproblems to improve the forecast accuracy and overall noise resistance of complex problems. This approach is called multitask learning algorithm.

Unlike single-task learning, multitask learning aims to address multielement and multidimensional learning problems. The input and output values of a system are considered multielements. The task number of multitask learning is generally determined by the dimensionality of the output value of the system. For example, if the system outputs a T-dimensional vector, then multitask learning learns T tasks at the same time.

Figure 1 shows the multitask learning algorithm process, which inherits the consistent aim of machine learning and where training dataset is directly constituted of the input and output values of the complex system. In the learning process, parallel learning is conducted on four learning tasks, and their results influence one another mutually. This process is the core idea of the multitask learning problem. Multitask learning can overcome the performance bottleneck of single-task learning and further improve the noise resistance of the algorithm and the accuracy of learning outcome.

Based on the security event evaluation data of a large-scale wireless sensor network, the primary algorithm of region similarity-based security event forecast for a large-scale network is as shown in Algorithm 1.

The basic idea of the algorithm is setting the target zone p as the main task and inputting measured index, task weight, regulation parameter h, and similarity positive threshold t_{positive} as the measurement network of all subregions. The output is the output set and errors verified by cross-validation. The algorithm first calculates the network

(1) Input: data for all sub-regions $\{(X_i^l, Y_i^l)\}$ $(l = 1, \ldots, m)$
(2) Parameters: h and $\lambda_1, \ldots, \lambda_m$, main task p and threshold t_{positive}
(3) Output: main task prediction data (X_i^p, Y_i^p), sim tasks S_j^p, and class loss CL_p
(4) Initialize: $u_l = 0$ $(l = 1, \ldots, m)$ $j = 1$, arbitrary $h \times p$ dimensional matrix Θ, $S_j^p = 0$
(5) iterate
(6) for $l = 1$ to m do
(7) If $l \sim= p$
(8) similarity $= S_{\text{Network}}((X_i^p, Y_i^p), (X_i^l, Y_i^l))$
(9) If similarity $\geq t_{\text{positive}}$
(10) S_j^p = similarity, $j = j + 1$
(11) With fixed Θ and $v_l = \Theta u_l$ approximately solve for $\widehat{w}_l$:
(12) $$\widehat{w}_l = \arg\min_{w_l} \left[\left(\frac{1}{n_l} \right) \sum_{i=1}^{n_l} L\left(w_l^T X_i^l + \left(v_l^T \Theta \right) X_i^l, Y_i^l \right) + \lambda_l \left\| w_l \right\|_2^2 \right]$$
(13) Let $u_l = w_l + \Theta^T v_l$
(14) end if
(15) end if
(16) end for
(17) Compute the singular value decomposition of $U = [\sqrt{\lambda_1} u_1 \cdots \sqrt{\lambda_m} u_m]$:
(18) $U = V_1 D V_2^T$ (with diagonals of D in descending order)
(19) Let the rows of Θ be the first h rows of V_1^T
(20) until converge
(21) using Θ through CV output (X_i^p, Y_i^p) and CL_p

Algorithm 1

similarities of target tasks in the network region and other subregions. If this similarity is larger than the set threshold, then its subregions are taken in learning tasks and iterate until convergence. Finally, the algorithm output is obtained through cross-validation.

2.3. Experimental Results and Analysis. A network topology with 66,072 routing nodes and 96,073 links is classified, and the connected subgraph with a topology size of approximately 1,000 nodes is selected as the experimental topology. A national topology with an actual measure of 66,072 routing nodes and 96,073 links is classified, and the connected subgraphs with a topology size of approximately 1,000 nodes are selected as experimental topology.

This topology is derived from Harbin Institute of Technology in the network topology nodes on the main stem distribution obtained from the measurements of actual topology data. The topology contains 66,072 routing nodes and 96,073 links. We divide the connected subgraphs with 1,000 (+50 or −50) nodes from the actual topology and experiment with 100 groups. The topological division algorithm metis [26] is used to divide the large-scale topology graph into approximately 1,000 nodes. The subtopology graph is connected.

In this topology, four worm attacks and Gnutella application are simulated, in which worm 1 attacks 159 groups, worm 2 attacks 30 groups, worm 3 attacks 30 groups, worm 4 attacks 30 groups, and Gnutella applies 89 groups.

The multitask learning algorithm adopts a linear classifier; therefore, the network damage characteristic value is obtained with weighted sum and discretization. The analyzed topology characteristic value is taken as input, and whether damage value can fall in this interval is taken as output. Each

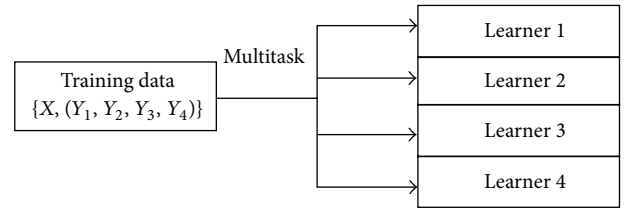


Figure 1: Diagram of the multitask learning algorithm.

interval is taken as one learning task, in which the damage value of the worm 1 attack is discretized into seven learning tasks, the damage values of the worm 2–worm 4 attacks are discretized into five learning tasks, and the damage value of the Gnutella application is discretized into four learning tasks. Thus, $159 * 7 + 30 * 5 * 3 + 89 * 4 = 1{,}919$ groups of samples are obtained.

The fifth learning task in the worm 1 attack is taken as the main task, and other tasks are taken as auxiliary tasks for learning. Cross-validation is also conducted. This process is mainly used in modeling application. In the given modeling samples, most samples are taken for modeling, whereas a small number of samples are established for forecasting. Their forecast error can be made to record their square sum. This process proceeds until all samples are forecasted once and only once. In the training process, fivefold cross-validation is performed; specifically, the dataset is divided into five, four for training and one for testing. The mean of five results is taken as the estimation of the algorithm precision. The final experimental results are as follows.

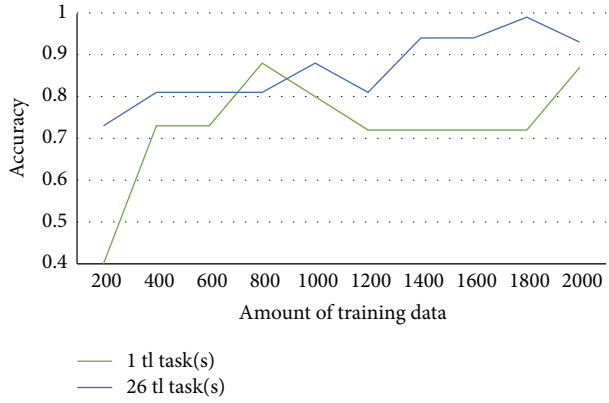


Figure 2: Mean square error of results.

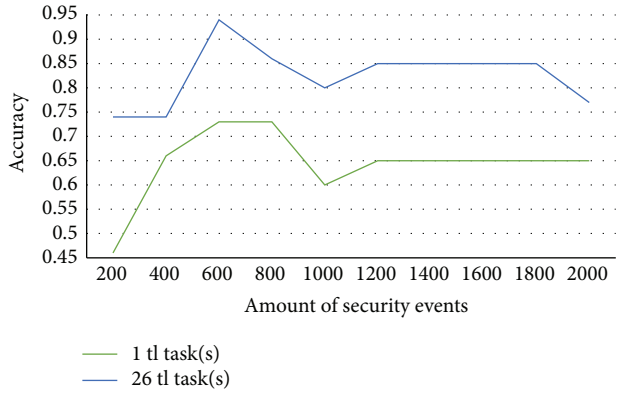


Figure 3: Comparison of the forecast results of unknown security events.

In the first experiment, one region is selected from the divided topology subdomain as the main zone. A total of 26 similar subregions are then selected as associated learning tasks to study the influences of increasing data amount of tasks on forecast accuracy. The training results of the experiment are shown in Figure 2.

Figure 2 presents that the accuracy of the task-joint-training results of 26 subregions is obviously higher than that of the single-training result of one subregion. In nearly 800 groups of training data, the forecast accuracy of the similarity region forecast is lower than that of single-task learning. The reason may be the cross-validation of the similarity region forecast. However, the data distribution characteristics of the randomly segmented training set and the validation set differ considerably, which decreases forecast accuracy. In general, the similarity region forecast algorithm has obvious effects on improving forecast accuracy.

Figure 3 indicates that for unknown security event forecasts the forecast accuracy of the classifier after training of the similarity forecast algorithm is higher and the effects are obviously better than those of the classifier after training of single-task learning.

In the second experiment, a region with a higher forecast error from 60 national subregions is selected as the target region. The task of the region is the main task. The data of other subregions form a set, and each task in the set is the forecast task of a region. The data in the set are then filtered. The sets that are completely dissimilar to the target subregion are removed, which results in a set that has remaining data of 47 regions.

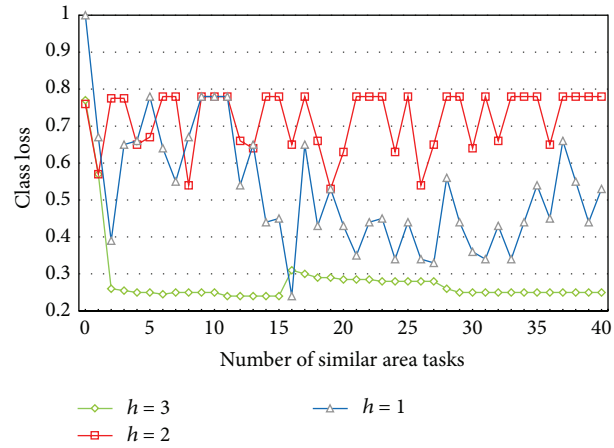


Figure 4: Regional similarity security event forecast error.

In the experiment, tasks in this set are gradually added in the forecast algorithm. The forecast error of adjusting different parameter h is shown in Figure 4.

Experimental results show that when the similarity network regions participating in learning are 3, the forecast error decreases to 0.26, which indicates that the tasks in the similarity region have good effects on improving forecast accuracy. However, when the tasks in the similarity region continue increasing, the improvement of forecast accuracy is not obvious. When the forecast tasks reach 17, error increases. Through checking, we find that the network environment of the 17th subregion is basically consistent with that of the target region, but its security events in the network are dense, thereby forming forecast noise of the target region. Therefore, the selection of proper similar regions requires both similarity measurement and analysis process with expert participation.

3. Security Event Forecast for Sensor Network in Case of Data Missing

When large-scale security spreads in each subregion, data (including throughput, packet loss probability, delay, and invasion index) in each subregion need to be collected to perform quantitative evaluations on network availability. However, given that many subregions have no IDS, the invasion index data of these regions may not be collected. In this case, part data of this region or data from other similarity subregions can be used to forecast the invasion index in data-missing regions and further obtain the quantitative evaluation values of network security events.

The PNSIUD method in case of data missing is adopted to solve the above problem. The data of the missing invasion index in the target region are forecasted in accordance with the basic idea of TSVM and the known data of similar regions

(including throughput, packet loss probability, delay, and invasion index).

3.1. TSVM Theory. The proposal of the transductive learning concept is established mainly from two facts. First, in the numerous actual applications of machine learning, tagged training samples usually cost significant manpower and material resources. Therefore, the tagged training sample from learning machine is usually limited, whereas numerous untagged samples are easy to obtain. The classification of missing invasion web pages in this study is a typical sample. In this application, manual or semiautomatic tagging of known web pages is usually trivial and dull. However, obtaining thousands of untagged web pages is easy. Considering that tagged training samples are few, the overall distribution characteristics of data cannot be well described, and the classifier that directly adopts tagged samples for learning usually has poor performance. Can the easily obtained untagged samples be used for improvement? Generating a problem on how to use a few tagged samples and numerous untagged samples to train a learning machine is natural.

Second, in the traditional inductive learning, a learning machine tries to summarize a discriminant function according to known tagged training samples. Consequently, the entire sample spatial distribution may have a low expected discriminant error. However, many actual problems only aim at the identification and classification of some specific samples, attempting to obtain the classification of a specific test set with a small error. If this specific test set is organically added in the design and training processes of the classifier, not only can this specific test set obtain good classification effects, but also the generalization performance of the original inductive learning algorithm can also be improved to a great extent. Furthermore, the shortcoming caused by few training data in the inductive learning algorithm can be made up for. This feature is the basic idea of transductive learning.

In transductive learning, a learning machine can use less tagged samples and more tagged samples in the training process for learning. An important feature of transductive learning lies in the sample distribution information of the test set being transferred from untagged samples to the final classifier in the learning process of mixed samples. Given numerous untagged samples, tagged samples can better describe the data characteristics of the entire sample space, such that the trained classifier can achieve improved generalization performance. Transductive learning is applied to different degrees of studies and applications in each field of mode recognition.

3.2. PNSIUD Method in Case of Data Missing. Based on the above theory, data without invasion index in the target region are called untagged data, whereas data with invasion index in the target region are called tagged data. Tagged data from this region or other similarity subregions are utilized to forecast the invasion index in the data-missing region. The principle and implementation of the PNSIUD algorithm are introduced below.

The training sample points with invasion indexes in the known similar subregions are

$$(x_1, y_1), \ldots, (x_n, y_n), \quad x_i \in R^m, \ y_i \in \{-1, +1\}. \tag{11}$$

The training points without invasion indexes in the target region are $x_1^*, x_2^*, x_3^*, \ldots, x_k^*$.

Under general linear inseparable conditions, the training process of the network security event forecast method in case of data missing can be described as the following optimization problems:

$$\begin{aligned}
\text{Minimize} \quad & \text{over}\,(y_1^*, \ldots, y_k^*, w, b, \xi_1, \ldots, \xi_n, \xi_1^*, \ldots, \xi_k^*) \\
& \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{n}\xi_i + C^*\sum_{j=1}^{k}\xi_j^* \\
\text{subject to} \quad & y_i\left[w \cdot x_i + b\right] \geq 1 - \xi_i, \quad \forall_{i=1}^{n} \\
& y_i\left[w \cdot x_j^* + b\right] \geq 1 - \xi_j^*, \quad \forall_{j=1}^{k} \\
& \xi_i \geq 0, \quad \forall_{i=1}^{n} \\
& \xi_j^* \geq 0, \quad \forall_{j=1}^{k},
\end{aligned} \tag{12}$$

where parameters C and C^* are named as regulation parameters. Their functions are similar to parameter C in formulae (12). Parameter C^* is called the influence factor of untagged samples in the training process, and $C^*\xi_j^*$ is called the influence item of untagged sample j in the objective function.

The training process of PNSIUD is the solving process of the above optimization problems. The training algorithm can be roughly divided into several steps.

Step 1. Parameters C and C^* are specified, and inductive learning is used to have an initial learning on tagged samples, thereby obtaining an initial classifier. According to one rule, positive tagged samples, N, in an untagged sample are specified.

Step 2. The initial classifier is used to classify untagged samples. According to the discriminant function output of each untagged sample N untagged sample with a maximum output can be temporarily assigned with a tag value, whereas others are endowed with a negative tag value. A temporary influence factor, C_{tmp}^*, is assigned.

Step 3. All samples are retrained. For new classifiers, a pair of tag signs in test samples with different tag values are exchanged according to certain rules. In this way, the objective function value in optimization problem (11) can decrease to the maximum. This step can be implemented repeatedly until no sample conforming to exchange conditions can be found.

Step 4. Temporary influence factor C_{tmp}^* is evenly increased, and Step 3 is repeated. When $C_{\text{tmp}}^* \geq C^*$, the algorithm ends, and results are considered as the output.

The tag sign exchange in Step 3 guarantees that the solution after exchange is better than that before exchange. The temporary influence factor in Step 4 from small to large gradually increases, and gradually increasing the influences of untagged samples on untagged samples can minimize the classification error of untagged samples. The specified C^* in Step 1 is limited. From the exit criteria in Step 4, the algorithm can end after limited-time implementation and output results.

3.3. Analysis of the Experimental Results of PNSIUD-Based Security Event Invasion Index Forecast. In this study, the sample data without invasion index in the target region are considered untagged, and they need to be tagged by the tagged sample training from this region or similar regions. Owing to the numerous untagged samples and the identification and forecast of specified target regions, PNSIUD has better forecast performance (tagging) than the traditional support vector machine (SVM) method.

The topological region with 66,072 routing nodes and 96,073 links is divided into several subregions, and the data of national security events that spread in each subregion in unit interval (there are many data without invasion index) are taken as experimental data. Tagged samples are taken as the auxiliary data for knowledge structure mining and further obtaining transfer knowledge structure. PNSIUD is utilized to forecast the region of untagged samples.

Below is the comparison of the forecast results between PNSIUD and SVM in one region. With an increase in untagged training samples, namely, samples without invasion index, from 100 to 1,000, the overall distribution characteristics of data can be better described because the PNSIUD method can utilize the knowledge of similar regions for learning. In this way, more samples lead to higher prediction accuracy. However, SVM only forecasts according to the known tagged samples and does not learn the knowledge of other similar regions. Consequently, its accuracy decreases, as shown in Figure 5.

Figure 6 is the forecast of those methods on data without invasion index in the same subregion. The forecast accuracy of PNSIUD is 64.48%, whereas that of SVM is 33.56%. Thus, PNSIUD has higher forecast accuracy than SVM.

The experiment indicates that when numerous untagged samples and a few tagged samples exist, the distribution characteristics of tagged samples cannot reflect the characteristics of the entire dataset. The accuracy of the SVM algorithm thus decreases. However, the PNSIUD method can learn from untagged samples and well reflect the data distribution characteristics of the entire dataset; hence, its accuracy can be improved.

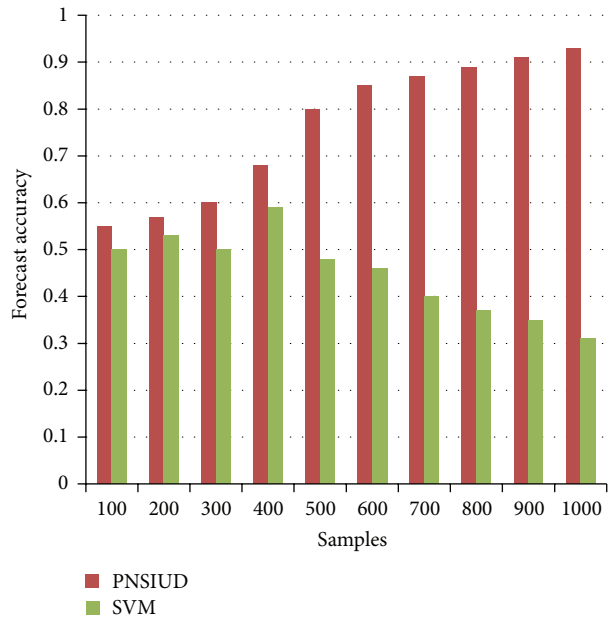


Figure 5: Comparison of forecast accuracy.

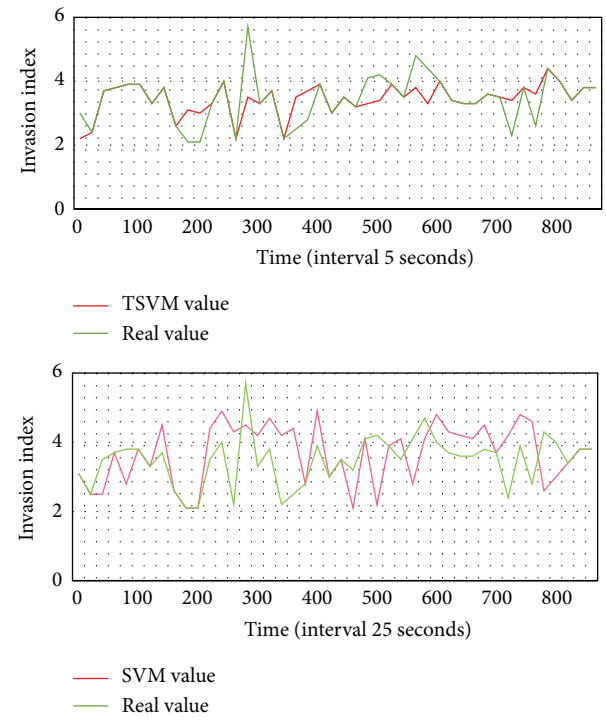


Figure 6: Network security event forecast in case of missing sensor network data.

4. Conclusion

This study conducts topology partitioning of a network from the characteristics of a large-scale sensor network. On this basis, this study proposes multitask learning-based similar region network security event forecast of wireless sensor networks. The experimental results show that for an unknown security event forecast the forecast accuracy and effects of classifier after the training of the similarity forecast algorithm are significantly better than those after the training of single-task learning. For data without invasion indexes in the same subregion, the forecast accuracy of PNSIUD is

64.48%, whereas that of SVM is 33.56%. Thus, PNSIUD has higher forecast accuracy than SVM.

Wireless sensor networks have strong dynamics and uncertainty, especially network topological changes. In this study, we take the topology and main factors in Table 1 as examples to illustrate how to forecast the missing attack data in the target region according to the known data in similar regions. Our work lacks some detailed factors in subregions. In the future, we will add more detailed factors into our research experiment, such as user characteristics and attack behavior. We will also introduce transfer learning into wireless sensor network security event forecast to improve the forecast adaptability.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work was supported in part by the National Basic Research Program of China under Grant no. G2011CB302605. This work is partially supported by the National Natural Science Foundation of China (NSFC) under grant nos. 61472108, 61173145.

References

[1] R. Ortalo, Y. Deswarte, and M. Kaâniche, "Experimenting with quantitative evaluation tools for monitoring operational security," *IEEE Transactions on Software Engineering*, vol. 25, no. 5, pp. 633–650, 1999.

[2] D.-G. Feng, Y. Zhang, and Y.-Q. Zhang, "Survey of information security risk assessment," *Journal-China Institute of Communications*, vol. 25, no. 7, pp. 10–18, 2004.

[3] L. Xiao and Z.-K. Dai, "Model of multilevel fuzzy comprehensive risk evaluation of information system," *Journal of Sichuan University (Engineering Science Edition)*, vol. 36, no. 5, pp. 98–102, 2004.

[4] L. Tao, "An immunity based network security risk estimation," *Science in China—Series F: Information Sciences*, vol. 48, no. 5, pp. 557–578, 2005.

[5] Y.-F. Wang, T. Li, X.-Q. Hu, and C. Song, "Real-time method of risk evaluation based on artificial immune system for network security," *Acta Electronica Sinica*, vol. 33, no. 5, pp. 945–949, 2005.

[6] Y. Wang, Z. L. Zeng, and X. Chen, "Quantitative method for risk assessment and management of information system," *Computer Engineering and Applicaiton*, vol. 22, pp. 8–10, 2005.

[7] Z. Yong, T. Xiaobin, and X. Hongsheng, "A novel approach to network security situation awareness based on multiperspective analysis," in *Proceedings of the International Conference on Computational Intelligence and Security*, pp. 768–772, IEEE, Harbin, China, December 2007.

[8] D. Wu, D.-G. Feng, Y.-F. Lian, and K. Chen, "Efficiency evaluation model of system security measures in the given vulnerabilities set," *Journal of Software*, vol. 23, no. 7, pp. 1880–1898, 2012.

[9] Q. Yan, "Information system security metrics and evaluation model," *Acta Electronica Sinica*, vol. 31, no. 9, pp. 1351–1355, 2003.

[10] K. Clark, S. Tyree, J. Dawkins, and J. Hale, "Qualitative and quantitative analytical techniques for network security assessment," in *Proceedings fron the 5th Annual IEEE System, Man and Cybernetics Information Assurance Workshop (SMC '04)*, pp. 321–328, June 2004.

[11] D.-M. Zhao, J.-H. Wang, W. U. Jing, and J.-F. Ma, "Using fuzzy logic and entropy theory to risk assessment of the information security," in *Proceedings of the International Conference on Machine Learning and Cybernetics (ICMLC '05)*, vol. 4, pp. 2448–2453, IEEE, Guangzhou, China, August 2005.

[12] D. Bilar, *Quantitative risk analysis of computer networks [Ph.D. thesis]*, Dartmouth College, Hanover, Germany, 2003.

[13] R. C. Cardoso and M. M. Freire, "Intelligent assessment of distributed security in TCP/IP networks," in *High Speed Networks and Multimedia Communications*, pp. 1092–1099, Springer, Berlin, Germany, 2004.

[14] X.-Z. Chen, Q.-H. Zheng, X.-H. Guan, and C.-G. Lin, "Quantitative hierarchical threat evaluation model for network security," *Journal of Software*, vol. 17, no. 4, pp. 885–897, 2006.

[15] H. Wang, J. Lai, L. Zhu, Y. Liang et al., "Survey of network situation awareness system," *Computer Science*, vol. 33, no. 10, pp. 5–10, 2006.

[16] H. Wang, J. Lai, M. Hu, and Y. Liang, "Research on key technologies for implementing network security situation awareness," *Geomatics and Information Science of Wuhan University*, vol. 33, no. 10, pp. 995–998, 2008.

[17] J.-H. Eom, S.-H. Park, Y.-J. Han, and T.-M. Chung, "Risk assessment method based on business process-oriented asset evaluation for information system security," in *Computational Science—ICCS 2007*, pp. 1024–1031, Springer, Berlin, Germany, 2007.

[18] W. Ren, *The intelligent study of networks security situation assessement [M.S. thesis]*, Shanghai Jiao Tong University, 2007.

[19] Z. Anwar, R. Shankesi, and R. H. Campbell, "Automatic security assessment of critical cyber-infrastructures," in *Proceedings of the IEEE International Conference on Dependable Systems and Networks With FTCS and DCC (DSN '08)*, pp. 366–375, Anchorage, Alaska, USA, June 2008.

[20] S. Kandula, R. Mahajan, P. Verkaik, S. Agarwal, J. Padhye, and P. Bahl, "Detailed diagnosis in enterprise networks," *SIGCOMM Computer Communication Review*, vol. 39, no. 4, pp. 243–254, 2009.

[21] Z.-H. Gong and Y. Zhuo, "Research on cyberspace situational awareness," *Journal of Software*, vol. 21, no. 7, pp. 1605–1619, 2010.

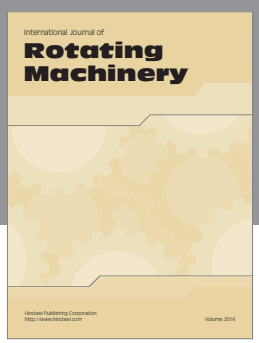
[22] R. Xi, X. Yun, S. Jin, and Y. Zhang, "Network threat assessment based on alert verification," in *Proceedings of the 12th International Conference on Parallel and Distributed Computing, Applications and Technologies (PDCAT '11)*, pp. 30–34, IEEE, Gwangju, South Korea, October 2011.

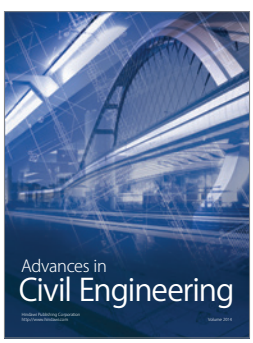
[23] A. B. Mohamed, N. B. Idris, and B. Shanmugum, "Alert correlation framework using a novel clustering approach," in *Proceedings of the IEEE International Conference on Computer & Information Science (ICCIS '12)*, vol. 1, pp. 403–408, Kuala Lumpur, Malaysia, June 2012.

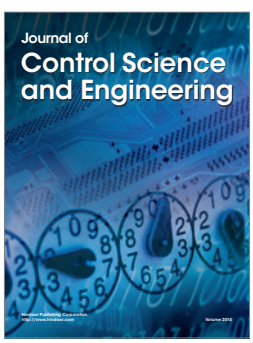
[24] Z. Tian, B. Wang, W. Zhang, J. Ye, and H. Zhang, "Network intrusion detection model based on context verification," *Journal of Computer Research and Development*, vol. 50, no. 3, pp. 498–508, 2013.

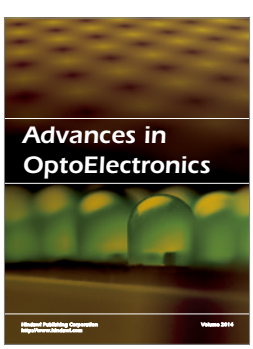
[25] W. Zhang, Y. Zhang, and T.-H. Kim, "Detecting bad information in mobile wireless networks based on the wireless application protocol," *Computing*, vol. 96, no. 9, pp. 855–874, 2014.

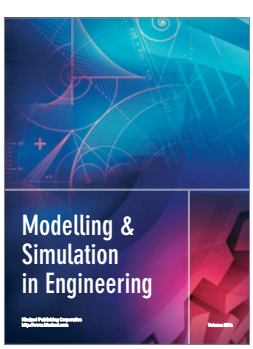
[26] G. Karypis and V. Kumar, "METIS—unstructured graph partitioning and sparse matrix ordering system," version 2.0., 1995.

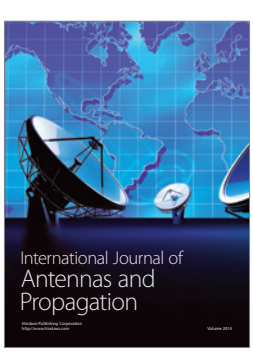


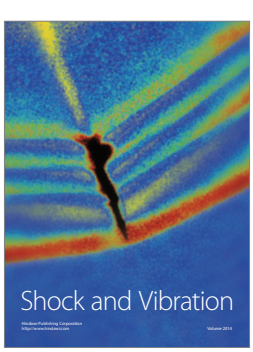


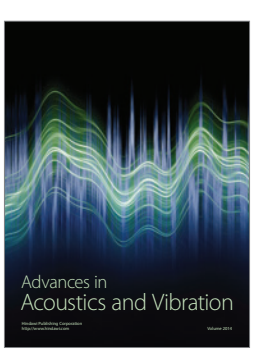












Hindawi

Submit your manuscripts at
http://www.hindawi.com