Ocean Wireless Sensor Network Location Method Based on Gradient Boosting Decision Tree

Liqun Liu
School of Mathematics and Computer Science, Guangdong Ocean University,
Zhanjiang, 524088, Guangdong, China

Bing Xu
School of Computer , Guangdong University of Petrochemical Technology,
Maoming, 525000, Guangdong, China

***Abstract*—**In order to solve the problem of low positioning accuracy of ocean wireless sensor networks (WSNs), a new positioning method based on gradient boosting decision tree (GBDT) is proposed in this paper. Firstly, the average positioning error is defined according to the hop count of the unknown node to the anchor node and the average connectivity degree of the network in this method. Next, the position coordinates of the unknown node are obtained. At the same time, the network topology is optimized. The free-space broadcasting model is used for power control, and signal interference and energy consumption are reduced consequently. A decision tree based on gradient is used to establish an ocean WSN positioning algorithm. Finally, the simulation experiment was out in MATLAB. The relationship among node communication distance, node density, link loss product and positioning error was analyzed. The results show that the method has good adaptability.

***Keywords*—**ocean wireless sensor network, gradient boosting decision tree, positioning error, node density

I. INTRODUCTION

ALONG with the deepening of ocean resources development and exploitation, positioning technology in ocean wireless sensor network (WSN) has begun to play an increasingly important role and become the research focus in the discipline of ocean environment[1, 2]. Ocean WSN consists of wireless sensors, network routers and signal nodes. Ocean environment is more complex than the continent, and conventional WSN has the defect of low positioning accuracy of nodes. Addressing this problem, many researches have been conducted at home and abroad and a series of models and algorithms have been established. Guo et al. [3] made a thorough exposition of underwater sensor network nodes and architecture, with the focus laid on research progress of underwater positioning technology. Guo et al. [4] designed the self-adaptive time synchronization algorithm for ocean parameters, which improved the adaptability of time synchronization algorithm to different ocean environments by adjusting the parameters. Simulation experiments showed that this algorithm was superior to conventional algorithms in both synchronization precision and energy efficiency. He[5] used time-of-arrival (TOA) ranging method for monitoring node positioning, and applied VBF routing protocol to WNS, which improved network robustness. Li et al. [6], using the Bayesian theory and continuous positioning algorithm, established new prior distribution and cost function from the positioning results of the previous moment. This method improved the positioning accuracy of the next moment and realized continuous positioning of moving sound source. In order to inhibit the impact of external strong vibration noise to measurement, Song et al. [7] proposed a vortex signal detection method based on multi-sensor information integration and performed data fusion by unscented Kalman filter algorithm, thereby lowering the interference. In literature report [8], infrared and microwave radar detection method was introduced. The atmospheric refraction over complex ocean environment was compared with that at the near-ocean level, and the differences in detection features in an optoelectronic system were obtained, which served as the basis for evaluation and analysis. For the strong reverberation environment of coaxial circular array, Yu et al. [9] proposed a moving multi-target detection method. The spatial information of the target was obtained by mode decomposition of linear frequency-modulated signals received and conventional beamforming algorithm. Literature [10] described a predictive node deployment algorithm for underwater acoustic sensor network. Integer linear programming theory, which resulted in a considerable increase in the positioning scope of the nodes. Zhang et al. [11] put forward an algorithm named CLIP and used in WSN location, simulation results indicate that the modified localization algorithm can greatly improve the localization accuracy of wireless sensor networks. In spite of the above researches, ocean WNS still has the defects of low positioning accuracy and overestimation.

Based on existing researches, new positioning method for ocean WNS based on gradient boosting decision tree (GBDT) is proposed in this paper [12-15]. This method provides the coordinates of unknown nodes and implements power control by the free-space broadcasting model. Then, this method is compared against conventional positioning algorithm. The contents of the article are organized as follows: section 1 reviews existing researches on ocean WNS; section 2 presents the performance indicators of ocean WNS; section 3 establishes the positioning algorithm for ocean WNS; section 4 is the simulation experiment; and section 5 is the concluding part.

II. PERFORMANCE PARAMETERS

The minimum hop counts of the known nodes x and y to the anchor node i are $d_{min}(x,i)$ and $d_{min}(y,i)$, respectively. $nbs(x)$ is the set of all neighborhood nodes of unknown node x. $d(x,i)$ is the hop count of the unknown node x to the anchor node i. $|nbs(x)|$ is the number of all neighborhood nodes of unknown node x. Then the hop count of the unknown node to the anchor node is calculated as follows:

$$d_{(x,i)} = \frac{\sum_{y\in nbs(x)} d_{\min(y,i)} + d_{\min(x,i)}}{|nbs(x)|+1} - 0.5 \quad (1)$$

Let R be the communication radius of the node, S the total area of the zone under monitoring, and n the total number of nodes. Then the average connectivity degree N of the network can be calculated by using formula (2):

$$N = \frac{n}{S}\pi R^2 \quad (2)$$

The average distance per hop D is given by

$$D = R(1+e^{-N} - \int_{-1}^{1} e^{-\frac{N}{\pi}(ar\cos t - t\sqrt{1-t^2})} dt) \quad (3)$$

Thus the distance from the unknown node to each anchor node is given by:

$$D_{(x,i)} = D\times d_{(x,i)} = R(1+e^{-N} - \int_{-1}^{1} e^{-\frac{N}{\pi}(ar\cos t - t\sqrt{1-t^2})} dt)(\frac{\sum_{y\in nbs(x)} d_{\min(y,i)} + d_{\min(x,i)}}{|nbs(x)|+1} - 0.5) \quad (4)$$

Let N be the number of unknown nodes, and the position of the unknown node i is estimated as (x_i,y_i). The actual position of the unknown node is $(\overline{x_i}, \overline{y_i})$. Then, the average positioning error is defined as :

$$e = \frac{\sum_{i=1}^{N}\sqrt{(x_i-\overline{x_i})^2+(y_i-\overline{y_i})^2}}{N\times R}\times 100\% \quad (5)$$

In a WNS, the anchor node first periodically sends the hop count packet and coordinates of the anchor node to the neighborhood nodes. The hop count is set to 0 during initialization. When the neighborhood nodes receive the information for the first time, they will update the minimum hop count of themselves to the anchor node. The minimum hop counts of the unknown node to each anchor node are recorded and updated. The hop count plus 1 and the coordinates of the corresponding anchor node are stored and forwarded to the adjacent node. Next, the distance between the nodes is estimated. That is, the records of the previous step are used to calculate the average distance per hop, including the coordinates of other anchor nodes calculated from each anchor node and the minimum hop counts to these nodes. Here, the hop count of the anchor node i at (x_i,y_i) to the anchor node j at (x_j,y_j) is expressed as h_{ij}. Thus for a total of n nodes, the corrected average distance per hop between the anchor nodes is

$$Hopsize = \frac{\sum_{i=1,j=1,i\neq j}^{n}\sqrt{(x_i-x_j)^2+(y_i-y_j)^2}}{\sum_{i=1,j=1,i\neq j}^{n} h_{ij}} \quad (6)$$

Literature [16] describes the regulated neighborhood distance method (RND), a new approach for estimating the distance between nodes and solving the obscurity of positioning in DV-Hop algorithm. This method uses the number of shared neighborhood nodes between the two neighborhood nodes to express the adjacency degree between them and to estimate the distance. Let N_i and N_j be the number of neighborhood nodes of node i and node j, respectively; and let the number of shared neighborhood nodes between node i and node j be n_{ij}. Thus the RND value can be calculated from the neighborhood nodes:

$$RND(i,j) = 1 - \frac{N_i+N_j}{2N_iN_j}\times n_{ij} \quad (7)$$

The shortest path RND_{min} between any two nodes can be found by using the Floyd–Warshall algorithm. Let the set of anchor nodes be X, and the Euclidean distance between the anchor node i and anchor node j be d_{ij}. Then, from the shortest path RND_{min}, the correction factor φ_{RND} for the unknown factor can be calculated:

$$\varphi_{RND} = \frac{\sum_{i\neq j,i,j\in X} d_{ij}}{\sum_{i\neq j,i,j\in X} RND_{\min}(i,j)} \quad (8)$$

The product of the correction factor and shortest path RND_{min} is obtained and used to derive the distance from the unknown node to the corresponding anchor node. To calculate the position using trilateration, it is necessary to know the distances from the unknown node to at least three anchor nodes.

As shown in Fig. 1, let the coordinates of unknown node D be (x,y). Then, the coordinates of nodes A, B and C are expressed as (x_a,y_a), (x_b,y_b) and (x_c,y_c), respectively; the distances from nodes A, B and C to node D be d_a, d_b and d_c, respectively. Thus, there is

$$\begin{cases}(x-x_a)^2+(y-y_a)^2 = d_a^2\\(x-x_b)^2+(y-y_b)^2 = d_b^2\\(x-x_c)^2+(y-y_c)^2 = d_c^2\end{cases} \quad (9)$$

From formula (9), the estimated coordinates of node D are obtained. On this basis, we can further know the position coordinates of the unknown node concerned:

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 2(x_a - x_c) & 2(y_a - y_c) \\ 2(x_b - x_c) & 2(y_b - y_c) \end{bmatrix}^{-1} \begin{bmatrix} x_a^2 - x_c^2 + y_a^2 - y_c^2 - d_a^2 + d_c^2 \\ x_b^2 - x_c^2 + y_b^2 - y_c^2 - d_b^2 + d_c^2 \end{bmatrix}$$
(10)

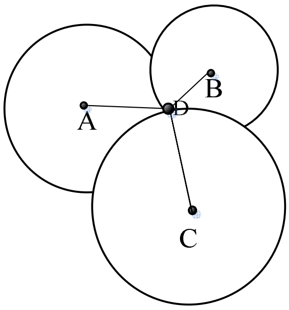

Fig. 1. Schematic for trilateration

III. MODIFIED ALGORITHM

Conventional positioning method in ocean WSN has the defects of high complexity of topological structure, signal interference and high energy consumption. Addressing these problems, we propose a positioning algorithm for ocean WSN based on GBDT. In order to reduce topological complexity, the topology of the Gabriel graph (GG) network is optimized. To reduce signal interference and energy consumption, free-space broadcasting model is used for power control, which further reduces the signal communication distance. Finally, the node mobility is learned by using GBDT. After learning the hidden connections in node mobility from the data, the high-accuracy node position will be recommended. The nodes are finally positioned using the formulae in section 2. The conditions for building GG, the topology connection graph based on graph theory, are as follows: for any two nodes i and j, if there is any node x other than nodes i and j that satisfies $d_{(i,x)}^2 + d_{(j,x)}^2 \le d_{(i,j)}^2$, then there is no direct connection between nodes i and j. Free-space broadcasting model is used for power control. Suppose that the transmission power of signals sent by the sending node at the source node is Pt, and Pr is the receiving power of the signals arriving at the receiving node. The transmitting gain and receiving gain are Gt and Gr, respectively, wavelength λ, link loss L, and communication distance between nodes d. Then, the broadcasting model is expressed as

$$P_r(d) = \frac{P_t G_t G_r \lambda^2}{(4\pi d)^2 L} \qquad (11)$$

Let the original sampling data of the WSN node be $X=[x_{ij}]_{n*m}$, i=1,2,3…,n, j=1,2,3,…,m, where n is the sample number, and m is the number of features (including the residual node energy, distance between node and base station and average connectivity degree of the network). The mean value of each feature $\bar{X} = \left[\bar{x}_j\right]_{1\times m}$ is calculated for the data matrix X. Demeaning is performed for X to obtain the normalized matrix $Y = \left[y_{ij}\right]_{n\times m}$:

$$Y = X - \bar{X} \qquad (12)$$

Covariance matrix *COV(Y)* is calculated for the normalized matrix $\bar{X}$:

$$COV(Y) = (n-1)^{-1} Y^T Y \qquad (13)$$

In order to get the eigenvalue λ_i and eigenvector η_i, the eigenvector and eigenvalue of the covariance matrix *COV(Y)* are analyzed:

$$\left|COV(Y) - \lambda E\right| = 0 \qquad (14)$$

where E is the $m*m$-order identity matrix. The eigenvalues are arranged in a decreasing order, i.e., $\lambda_1 \ge \lambda_2 \ge \lambda_3 \ge \cdots \ge \lambda_m$. Correspondingly, the eigenvectors are $\eta_1,\eta_2,\eta_3,\ldots,\eta_m$. Variance contribution ξ_i is calculated for each eigenvalue λ_i, i=1,2,3…,n:

$$\xi_i = \left|\frac{\lambda_i}{\sum_{i=1}^{n} \lambda_i}\right| \times 100\% \qquad (15)$$

Let the threshold for cumulative variance contribution be Q. k pivot elements are chosen to describe the majority of the feature information in the original samples:

$$\sum_{i=1}^{k} \xi_i \ge Q \qquad (16)$$

From the above formula, eigenmatrix after $R^m \to R^k$ eigenspace transformation is obtained. This is the matrix consisting of eigenvectors corresponding to k pivot elements. Since there is data loss during false information classification and recognition, logarithmic loss function is used:

$$L(y, F(x)) = 2\sum_{i=1}^{f} \log(1 + \exp(-2 y_i p_i)) \qquad (17)$$

where n is the number of samples; f is the number of features recognized in the false information; y_i is the actual label of the samples; p_i is the predicted label of the samples.

Based on the above formula and description in section 2, positioning algorithm in ocean WSN based on GBDT is derived:

Step 1 Initialize the model and original sampling data $X=[x_{ij}]_{n*m}$ at WSN nodes;

Step 2 Select qualified nodes according to formula $d_{(i,x)}^2 + d_{(j,x)}^2 \ge d_{(i,j)}^2$ to establish the topological structure of WSN;

Step 3 Extract feature information for WSN nodes using formulae (12)-(16);

Step 4 Let the threshold for cumulative variance contribution be Q. Select k pivot elements that satisfy $\sum_{i=1}^{k} \xi_i \ge Q$ to describe the majority of feature information in the original samples; otherwise, return to Step 3;

Step 5 Train the GBDT model:

Step5.1: Calculate constant α that makes the loss function minimal using formula $F_0(x) = \arg_\alpha \min \sum_{i=1}^{n} L(y_i, \alpha)$.

Start iterations for M generations from m=1. Build the model successively by performing descent along the direction of the gradient of the loss function in the previous step:

$$F_0(x) = \arg_\alpha \min \sum_{i=1}^{n} L(y_i, \alpha)$$

Step5.1.1 Calculate the value of the negative gradient of loss function in the current model, as the estimate value of residual r_{im}:

$$r_{im} = -\left[\partial L(y_i, F(x_i)) / \partial F(x_i)\right]_{F_m(x)=F_{m-1}(x)}, i = 1, 2, \cdots, n \ ;$$

Step5.1.2 Takes the residual in Step 5.1.1 as input and calculate the leaf node of the regression tree $R_{j,m}$(j=1,2,…,J) as well as the optimal step length α_{jm} along gradient descent:

$\alpha_{jm} = \arg_\alpha \min \sum_{x_i \in R_{j,m}} L(y_i, F_{m-1}(x) + \alpha)$; update the model

$F_m(x)$: $F_m(x) = F_{m-1}(x) + \sum_{j=1}^{J} \alpha_{jm} I; x \in R_{jm}$;

Step5.2 After the iterations are over, model $F_M(x)$:

$F_M(x) = \sum_{m=1}^{M} \sum_{j=1}^{J} \alpha_{jm} I; x \in R_{jm}$ is derived.

Step6 Analyze K pivot elements in Step 4 using model $F_M(x)$ to get the node position information;

Step7 Estimate the distance from the unknown node to the corresponding anchor node based on the product of correction factor $\varphi_{RND} = \frac{\sum_{i \neq j, i, j \in X} d_{ij}}{\sum_{i \neq j, i, j \in X} RND_{\min}(i, j)}$ and shortest path RND_{min};

Step8 Repeat Step 6 and Step 7, until there are no unknown nodes to be positioned;

Step9 Perform positioning calculation by trilateration. The algorithm is over.

IV. Simulation Experiment

To verify the applicability of the described algorithm, simulation experiment was conducted for the ocean WSN in MATLAB, with the scenario of extensive monitoring of regional ocean. WSN nodes were not deployed one by one, but randomly distributed. The environmental factors for the simulation were configured as follows: 400 nodes were deployed at the most, and were randomly distributed in a square of 1000m*1000m. There were 100 anchor nodes. The transmission radius was 100m for all nodes. The GBDT model parameters were optimized and the modified gradient was initialized: the maximum number of iterations T_{max} was set to 100, learning rate 0.1, leaf node depth 3. The loss function was defined by polynomial abnormal loss function.

The impact of the communicating node on the surrounding nodes was first simulated. Fig. 2 shows the comparison of the impact of the communicating node on the surrounding nodes without power control under original topological structure and with power control under GG topology. Under the latter situation, power control effectively reduced the distance and scope of signal interference, thereby reducing the impact of communication on the surrounding nodes. This was much better than without power control under original topological structure. Obviously, power control based on GG could optimize topological control and signal interference control. The feasibility of this technology was also verified by comparing against the experimental results.

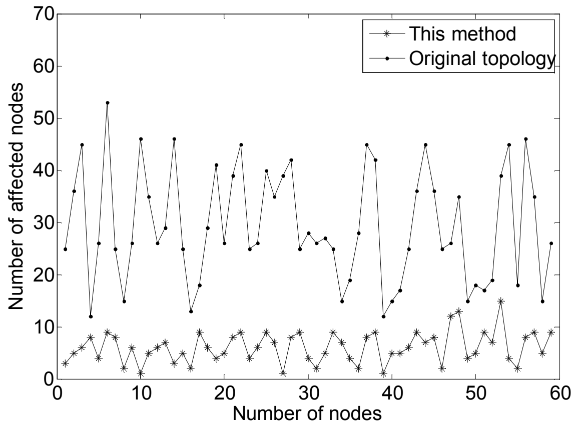

Fig. 2. Comparison of impact of the communicating node on the surrounding nodes

Fig. 3 shows the influence curves of number of unknown nodes on the node communication distance and link loss product using different algorithms. It was assumed that the communication radius was equal for all nodes (100m) and that the number of nodes varied from 75 to 280. It could be observed that as the number of nodes increased rapidly, the product of error and energy consumption increased for both algorithms. The increase was more significantly for DV-HOP algorithm. One standard for a good balance between positioning accuracy and energy consumption was the small increase, and our proposed algorithm had a very small increase.

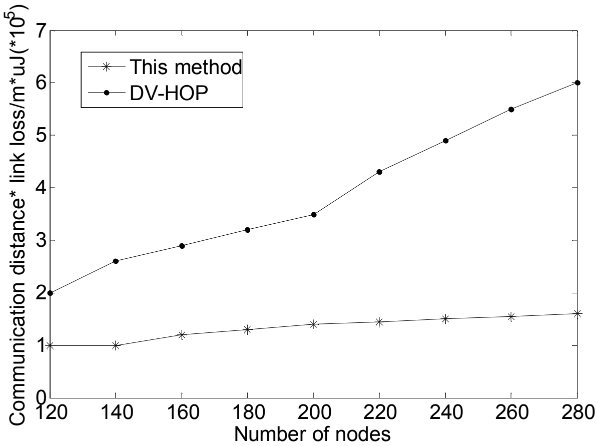

Fig. 3. Influence curves of number of unknown nodes on node communication distance and link loss product using different algorithms.

ISSN: 2692-5079

The connectivity degree of network is mainly influenced by node density. Therefore, it is necessary to examine the impact of node density on algorithm performance. Suppose that the ranging error was 0.05 and that the proportion of anchor nodes was 25%. Fig. 4 shows the relationship between Positioning error distance and node density. It can be seen that as the node density increased, the Positioning error distance first decreased and then increased. When the node density was below 9, the positioning error in the network decreased slowly; when the node density was 9-15, Positioning error distance increased; when the node density was above 15, Positioning error distance increased rapidly. Therefore, by setting node density as 9 we could obtain the lowest Positioning error distance and the optimal algorithm performance.

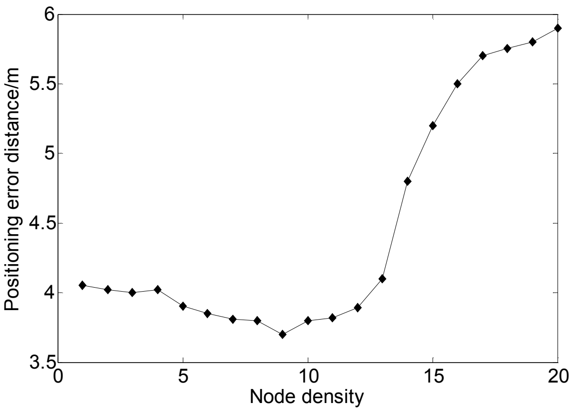

Fig. 4 Influence of node density on Positioning error distance

After several simulations, the Positioning error distance of our algorithm and DV-HOP were calculated, as shown in Fig. 5. Our algorithm outperformed DV-HOP not only in accuracy, but also in convergence property. In Fig. 5, the average Positioning error distance of the two methods was 4.5m and 9m, respectively. The positioning accuracy of our method improved by 50% as compared with DV-HOP. Moreover, the fluctuation of accuracy was much smaller with our method. Therefore, our method had higher positioning accuracy and convergence property than DV-HOP.

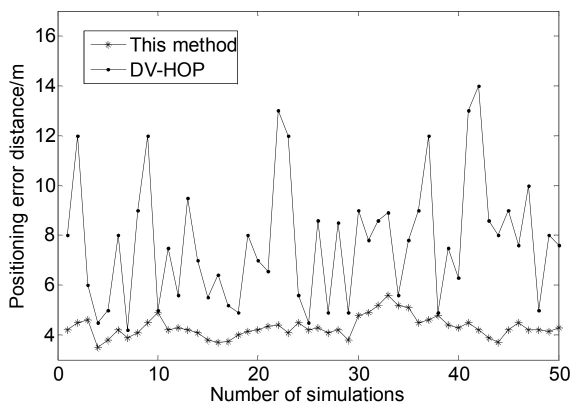

Fig. 5. Comparison of Positioning error distance using two algorithms

V. CONCLUSION

To solve the problem of low positioning accuracy of ocean WSNs, a new positioning method based on GBDT is proposed. Firstly, the average positioning error relative to the actual position of the unknown node is estimated according to the hop count of the unknown node to the anchor node and the distance between these two nodes. Next, the network topology is optimized by GG to reduce topological complexity. The communication distance is reduced in this way. The free-space broadcasting model is used for power control, and as a result the signal interference and energy consumption are reduced. Finally, the node mobility is learned by using GBDT. The hidden connections in node mobility are learnt from data and the high-accuracy node position is recommended. To verify this algorithm, simulation experiment was carried out with ocean WSN in MATLAB, and a comparison was made with DV-HOP algorithm in the following aspects: influence of the communicating node on the surrounding nodes, influence of number of unknown nodes on the node communication distance and link loss product, and influence of node density on positioning error. The results showed that our method outperformed the DV-HOP algorithm. For future studies, integration of other positioning methods may be considered, so as to further improve the positioning accuracy and reduce the computational time of the system.

REFERENCES

[1] Guo Y, Liu Y T. “Mobile counteracted time synchronization for underwater sensor networks”, *Information Computing and Applications*, vol. 307, no.1, pp. 515-522,2012

[2] Liu J, Zhou Z, Zheng P, Cui J H. “Mobi-Sync: efficient time synchronization for mobile underwater sensor networks”, *IEEE Transactions on Parallel and Distributed Systems*, vol. 24, no.2, pp. 406-416,2013

[3] Guo Zhongwen, Luo Hanjiang, Hong Feng. “Current Progress and Research Issues in Underwater Sensor Networks”, *Journal of Computer Research and Development*, vol. 3, no.47, pp. 377-389,2010

[4] GUO Ying, WANG Jin-xin. “Ocean parameter adaptive time synchronization algorithm for ocean sensor network”,*Journal on Communications*. vol. 3, no.47, pp.2017237-12017237-6,2017

[5] HE Xie-fei. “Research on a new network routing algorithm for ocean monitoring node localization”,*SHIP SCIENCE AND TECHNOLOGY*. vol. 5, no.39, pp. 187-189,2017

[6] Li Qian-Qian, Yang Fan-Lin, Zhang Kai. “Moving source parameter estimation in an uncertain environment”, *Acta Phys. Sin*. vol. 16, no.65, pp. 164304,2016

[7] Songkaicheng,Zengyao,Yelingyun. “Vortex signal processing method based on multi-sensor information fusion”,J*ournal of Zhejiang University (Engineering Science)*, vol. 7, no.50, pp.1307-1312,2016

[8] LI Yun-bo,ZHANG Yong-gang, WANG hua. “Influence of marine abnormal refraction environmental on infrared and microwave transmission”, *Infrared and Laser Engineering*, vol. 37, pp.254-259,2008

[9] Yuge, Piao shenchun. “Multiple targets detection in strong revereration environments”, *J.Huazhong Univ. of Sci. & Tech.(Natural Science Edition)* , vol. 3, no.45, pp. 89-93,2017

[10] Chen Qiuli,He Ming, Wang Yan. “Localization algorithm of underwater acoustic sensor network oriented to marine monitoring”,*Computer Science*, vol. 2, no.42, pp. 29-32,2015

[11] Zhang Xue,Wang Weigang,Tian Longbin. “Modified Localization Algorithm for WSNs based on RSSI”,*Communjcations Technology*, vol.

5, no.50, pp. 968-972,2017

[12] Yang Niya,Peng Tao, Liu Lu. "Link prediction method based on clustering and decision tree", *Journal of computer research and development*, vol. 54, no.8, pp. 1795-1803,2017

[13] WANG He-Peng, WANG Hong-Zhi, LI Jian-Zhong, GAO Hong. "Algorithms for Accurate Decision Tree Generation on Inconsistent Data", *Journal of Software*, vol. 28, no.11, pp. 2814-2824,2017

[14] Guo Taisheng,Zhang Xinchang,Liang Zhiyu. "Research on change information recognition method of vector data based on neural network decision tree", *Acta geodaetica et cartographica sinica*, vol. 42, no.6, pp. 937-944,2013

[15] Ji Yimu,Zhang Yongpan, Lang Xianbo. "Parallel of decision tree classification algorithm for stream data",*Journal of computer research and development*, vol. 54, no.9, pp. 1945-1957,2017

[16] Wu G, Wang S, Wang B, Dong Y, Yan S. "A novel range-free localization based on regulated neighborhood distance for wireless ad hoc and sensor networks", *Computer Networks*, vol. 56, pp. 3581-3593,2012