SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

January 2018, Vol. 61 012106:1–012106:19
doi: 10.1007/s11432-016-9071-5

Mission evaluation: expert evaluation system for large-scale combat tasks of the weapon system of systems

Jianfei DING[1,2], Guangya SI[1*], Jun MA[1], Yanzheng WANG[1] & Zhe WANG[3]

[1]*Department of Information Operation and Command Training, National Defence University of PLA, Beijing* 100091*, China;*
[2]*Graduate School, National Defence University of PLA, Beijing* 100091*, China;*
[3]*Department of Physiology, Hebei Medical University, Shijiazhuang* 050017*, China*

Received 29 November 2016/Accepted 16 March 2017/Published online 19 July 2017

Abstract Mission evaluation is a new requirement for capability evaluation of the weapon system of systems (WSOS) in the era of big data, and is based on evaluating large-scale tasks with similar attributes. The use of traditional methods by military experts to evaluate large scale tasks incurs significant time cost and results in low accuracy, and is caused by a variety of factors that cause confusion. Therefore, we developed a system to assist military personnel in improving the efficiency of mission evaluation; the main innovations of our work include the qualitative and quantitative visualization of complex information is realized in a three-pane interface. We also realize the iterative and interactive evaluation modes of large-scale tasks by using the active learning method; moreover, the overall display of large-scale task evaluation results is realized using statistical graphics. In practical application, the system not only improves the users' efficiency and accuracy scores, but also helps to achieve the recognition evaluation for the overall scoring results.

Keywords mission evaluation, large-scale tasks, weapon system of systems (WSOS), visual interactive, active learning

Citation Ding J F, Si G Y, Ma J, et al. Mission evaluation: expert evaluation system for large-scale combat tasks of the weapon system of systems. Sci China Inf Sci, 2018, 61(1): 012106, doi: 10.1007/s11432-016-9071-5

1 Introduction

In system engineering, war is a special social system known as an open complex giant system (OCGS). From his research on three giant complex systems (social systems, human systems, and geographical information systems), China's famous scientist Tsien Hsue-Shen drew the conclusion that the only effective method for studying OCGSs is a combination of qualitative and quantitative methods. Modern war is the confrontation between combat systems of systems, which are substantially WSOSs when not considering human factors. A WSOS is a higher level of weapons system composed of various other weapon systems whose functions contact and interact to complete certain missions under the guidance of a certain strategy. Obviously, WSOS is a complex system, and operation actions and tasks are always affected by various complex factors. Therefore, the performance of WSOS cannot be evaluated with a simple quantitative calculation; instead, it often requires comprehensive qualitative and quantitative analyses by experts. In the traditional evaluation of WSOS for a specific task, military experts simply

* Corresponding author (email: sgy863@sina.com)

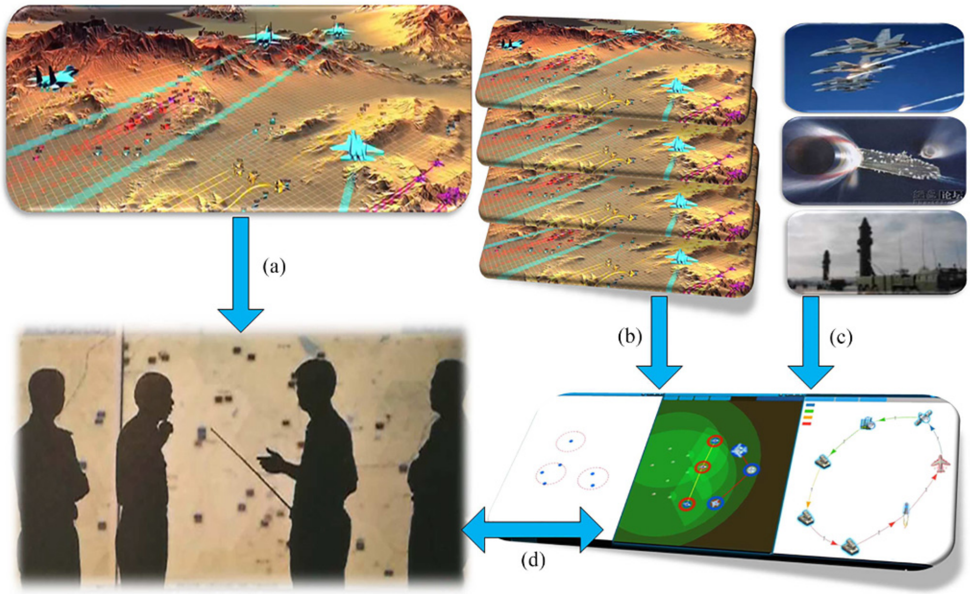

Figure 1 From task evaluation to mission evaluation. (a) Traditional patterns of task evaluation; (b) mission evaluation based on large scale tasks; (c) mission packages including multiple missions; (d) visual interaction and iteration.

draw a conclusion from one to several simulation tasks, as shown in Figure 1(a). However, the results are often partial and cannot reflect the integrity and the uncertainty of WSOS capabilities. The current evaluation is more concerned with a specific mission, which is a set of large-scale tasks under a general goal of WSOS, as shown in Figure 1(b). The completion degree of the mission can comprehensively and objectively reflect the capability of the WSOS, but there will be more challenges. The first of the two main problems to solve is the scarcity of high-level military experts. In military practice, the evaluation criteria cannot be simple quantification, and mission evaluation by experienced commanders and high-level military experts, combining both qualitative and quantitative methods, is often required. An important secondary facet is improving the accuracy of expert evaluation, as the complex relationship of task attributes will cause confusion of evaluation criteria as the number of tasks increases.

Based on the principle of combining qualitative and quantitative analysis, we use a novel approach to integrate visual, interactive, and active learning techniques into the evaluation of operational tasks, and develop an expert mission evaluation system. The system can improve the quality and efficiency of large-scale task evaluation in two aspects: the informational expression of tasks and the intelligent learning of experts' scoring rules. Compared to the known military evaluation methods and systems, the system in this paper is a new mode for expert evaluation. Specifically, it has the following contributions:

(i) A three-pane interface: battlefield task information visualization, attribute space clustering, and observe-orient-decide-action (OODA) combat interactive network visualization.

(ii) The iterative scoring method based on active learning: obtaining prediction evaluations of large-scale tasks from a few tasks scored by experts.

(iii) Feature information display of evaluation results: demonstrating the features of evaluation results from multiple perspectives based on statistical indicators and graphics.

Based on the above methods and modes, military experts can efficiently evaluate all kinds of missions (Figure 1(c)) by way of visual interaction and iteration relying on the system as shown in Figure 1(d), and the evaluation results of these tasks can be analyzed intuitively from multiple angles. Compared with traditional methods, our system not only considers the quantitative factors of the task itself, but also considers the qualitative factors of the whole, that is, a mixed mode. These features ensure the practicality and accuracy of our approach.

2 Related work

Since the beginning of the century, some research groups in China have studied and constructed the hall for workshop of meta synthetic engineering (HWME) to solve practical field problems for the expert evaluation problem of OCGS according to the comprehensive integration method [1] proposed by Tsien Hsueshen. It is essentially a highly intelligent human-machine interaction system comprised of expert groups, information data, and computer technology. Dai of the Institute of Automation, Chinese Academy of Sciences, with his research team, designed the cyber space for workshop of meta synthetic engineering (CWME) based on cyber space; their method involved the construction of an expert discussion environment without time and geographical restrictions, and was used in macroeconomic decision support [2–4]. The National Defense University team led by Hu and Si put forward an integrated idea of on demand services, and constructed a virtual simulation environment servicing national defense strategy decision making [5, 6]. Zhang and his team at Shanghai Jiaotong University studied the visual performance of experts' views in group discussion [7, 8]. Xiong at Hubei University of Technology conducted visualization of reasoning processes in group discussion [9]. The above systems focus mainly on facilitating the collaboration of experts and making their respective views accessible.

2.1 Game analysis

Ball games provide a good reference for the natural similarity they share with WSOS regarding characteristics of team cooperation and confrontation. Although there are differences between the specific data forms of weapon systems and the ball game data, the characteristics of their data as a whole are similar. For example, both include statistical data, track data, and movement of time and space data. Current research on these games not only focuses on the quick and accurate understanding the game information, but also analyzes the player or team performance in the game; i.e., player efficiency, coach substitution strategy, and pause strategy. The analysis of basketball games: Ref. [10] measured the efficiency of the players and the team, Ref. [11] analyzed the characteristics of the game data, and Ref. [12] analyzed the characteristics of players. The analysis of football games: Refs. [13, 14] showed the differences of performance between the players using visual methods, Ref. [15, 16] performed a visual analysis of temporal and spatial characteristics on the team's overall performance, Ref. [17] searched and visualized the characteristics of important moments of games and events using a classifier. In the above game researches, visualization and machine learning methods can help professionals to conduct in-depth analysis and evaluation of games and players, which is similar to the combat analysis of military experts. The completion degree of tasks in the game can be evaluated by scores; however, the evaluation of tasks in war must consider more complex factors involving the tasks themselves and their overall impact on real combat. Therefore, the comprehensive judgments of high-level experts are necessary to military task evaluations. In order to make up for the shortage of experts, reduce the workload of high-level experts, and improve the accuracy of machine learning on expert evaluation rules, we simultaneously study active learning methods.

2.2 Active learning method

The active learning method was originally proposed by professor Angluin at Yale University, and published in her paper Queries and Concept Learning [18]. To overcome deficiencies in semi-supervised learning and achieve cost-effective active learning, this study not only considers marked data as prior knowledge to train, but also chooses optimum unmarked data to query, thus stimulating the human learning process of ask and answer. Active learning has drawn extensive attention and has developed rapidly as it enables high performance classification of large scale samples using fewer training samples. A survey by Tomanek and Olsson [19] showed that 90.7% of researchers used the active learning algorithm to achieve desired results in their projects; results of a separate survey show that a few multinational IT companies such as Google, IBM, Microsoft, and Siemens have introduced active learning algorithms to improve the learning effect in current projects [20, 21]. In the academic field of machine learning, the active learning method has always been a hot issue in important academic conferences and leading international journals in recent years.

Active learning has been widely used in many fields; for example, network data classification [22–26], text categorization [27, 28], image retrieval [29], emotion classification [30], video retrieval [31], stream data mining [32], information retrieval [33], image segmentation [34, 35], three-dimensional model construction [36, 37] and personalized recommendation [38–40]. Through its application in different fields, it is known that to obtain good learning results, we must make full use of the inherent characteristics of field data and a good visual interactive environment to help field experts make the best choices. For example, in the field of network data classification, ref. [22] used the center of a graph and community discovery to study the relationship between network data, and measured the information content of samples from various angles using multiple strategies. Existing literature [24] determined that the influence of nodes was an important measurement criterion by using dependence between nodes. In the field of text categorization, the relationship between texts was displayed by way of the network topology, and a visual interactive environment for active learning was realized [28]. In the field of image segmentation, according to the characteristics of an image, Top et al. [35] considered the image pixel point gradient, curvature, and the uncertainty of the partition results as the comprehensive basis for the query. In the field of 3D model construction, Gao et al. [36] studied global and local features by using the decomposition feature of the model, and mapped the 3D model onto 2D plane to assist users when making choices.

By summarizing the above methods, we find that the combination of active learning and visualization can effectively help experts in the field to work. Specifically, it can improve the accuracy of expert evaluation and realize automatic scoring by learning expert evaluation rules.

3 Workflow of the system

To more clearly describe our system, we provide an overview of the entire process of system usage.

Step 1. According to the specific mission of a certain type of WSOS, the experiment personnel carry out multiple simulation experiments relying on the experiment platform of the simulation, and the simulation experiment data are thereby produced.

Step 2. The mission evaluation system extracts data related to task evaluation, and constructs the task set space where these tasks have not been evaluated.

Step 3. These task data are projected into a three-pane visualization interface using three visualization methods: the similarity space visualization method based on task attribute clustering, task information visualization method based on the battlefield environment, and overall impact visualization based on an OODA interactive network.

Step 4. The first scored tasks are selected independently by an expert based on the three-pane interactive interface.

Step 5. According to evaluation information, the classification training of the task set space would be taken by the machine learning model. The active learning model then selects the tasks most needed to be queried, and provides them to the experts for scoring.

Step 6. The evaluation results of all tasks are displayed by the statistical graphical interface as the evaluation basis of the WSOS mission.

The specific process is shown in Figure 2.

4 System implementation

4.1 Task definition

Here, we briefly define the concepts and mathematical symbols involved in this paper. It is assumed there are a total of n tasks in the task instance set $\mathrm{TD} = \{x_1, \dots, x_n\}$. The first scored task set $L = \{(x_1, y_1), \dots, (x_l, y_l)\}$ is obtained after experts scored l tasks, and $y_i \in Y = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$ is

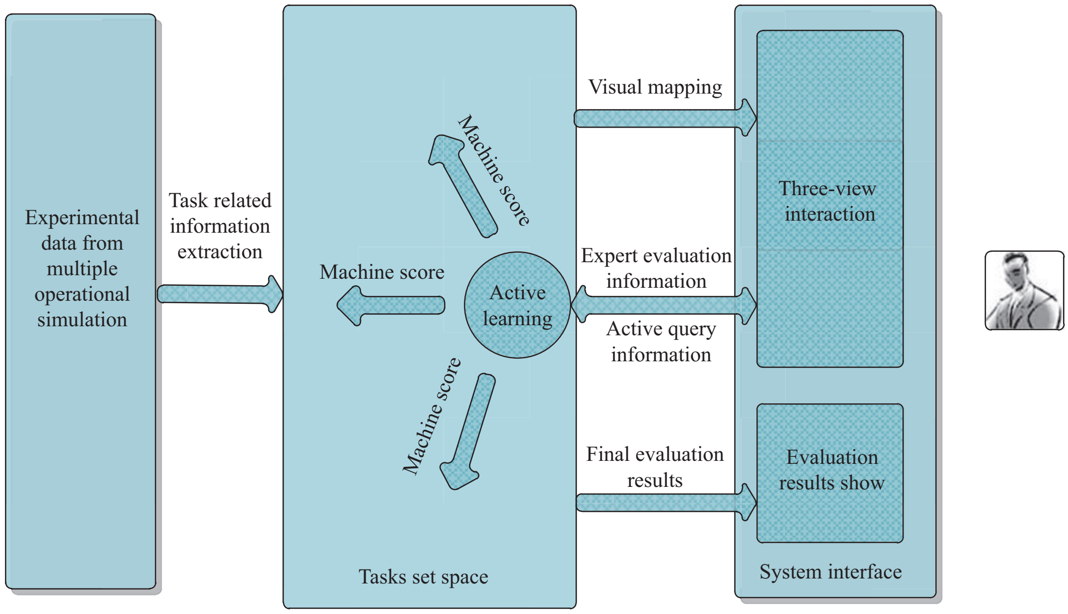

Figure 2 Work flow chart of the system.

the value of the task score; other non-scoring tasks are defined as $U = \{x_{l+1}, \ldots, x_{l+u}\}$, where $n = l + u$ and $l \ll u$. Each task instance x_i has several task attributes x_{ia}; the number and content of the specific attributes are determined by the specific evaluation task. Taking the air defense early warning mission in this paper as an example, task attributes include warning type, target orientation, target height, and target type.

4.2 Three-view interactive interface

Whether in ancient or modern war, visualization is an important tool for military personnel to analyze battlefield information; therefore, the visual interface is a very important part of our mission evaluation system. Owing to the nature of evaluating a combat mission, our system not only considers the completion of the task itself, but also considers the impact of the entire operational process; moreover, it cannot be divorced from the real operational environment. Based on the above three points, we designed the interactive interface of the three views to realize the combination of qualitative and quantitative visual analysis methods. To improve the efficiency of the analysis, the visual content of the three-view interface can be dynamic depending on the selections of the operator, and the relevant correlations are shown. The structure is shown in Figure 3, and the interface is shown in Figure 4. Thus, to provide a more detailed description of this part, we take the air defense and early warning tasks as examples.

4.2.1 *Visualization of the battlefield environment based on task information*

To assess the capability of an air defense SOS for the early warning of air targets, it is imperative to simulate a large number of diverse warning tasks. For traditional task oriented evaluation, military personnel can effectively evaluate the tasks in the original battlefield environment visualization, as the number of tasks is low. However, for mission oriented evaluation, what and how to display battlefield information to help military experts to the greatest possible extent becomes the challenge of our work. First, we visually integrate the information of the WSOS related to a combat mission into the battlefield environment. Second, we identify the space position of the task in the form of the target. Finally, we distinguish the target signature based on the main attributes of the specific combat tasks. The battlefield environment visualization of an air defense WSOS for an early warning mission is shown in Figure 4(b). The information of the WSOS is as follows: the green and circular sectors represent the warning range of the air defense WSOS. Task space position: six aircraft icons in the air represent the space position of

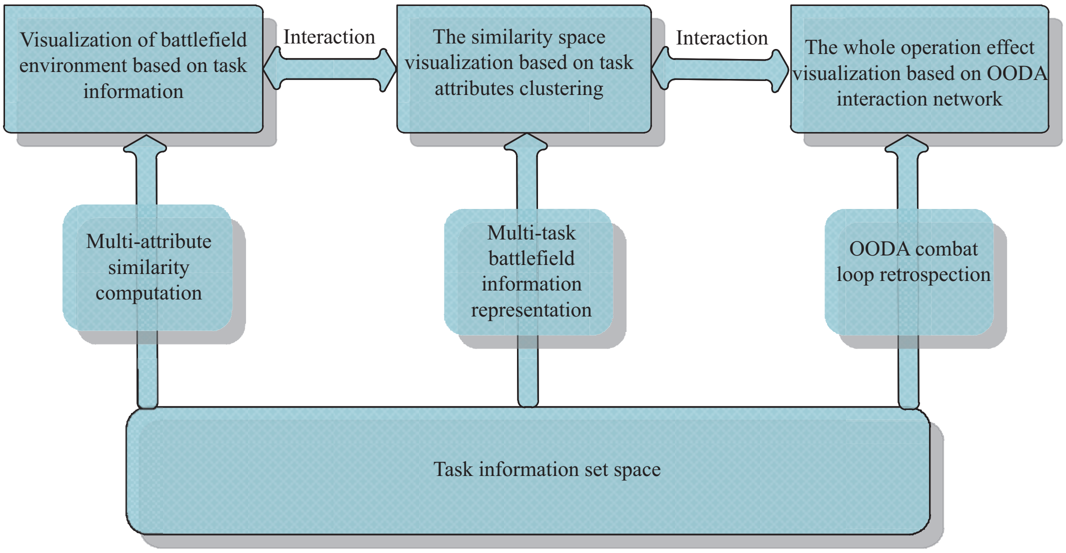

Figure 3 Structure graph of three-view method.

the six early warning tasks. Task attribute information: different aircraft icons represent different kinds of targets; the outer ring represents the identification of friend or foe (IFF) task, the inner ring represents the non-cooperative target recognition (NCTR) task, non-ring represents the detect (DET) task, and the color of ring represents the side of the target. The connection lines between targets represent a high range of targets (e.g., yellow represents a higher elevation, red represents a lower elevation). Based on the above information display when superimposed, experts can make quick evaluations when only one single attribute is different. For example, if all task attributes are same except for target type, then the task score will be higher when the aircraft is more advanced. However, it is difficult to make a clear judgment when there are a lot of tasks with multiple different attribute values. Thus, we provide the second visualization interface.

4.2.2 *The similarity space visualization based on task attributes clustering*

By calculating of the similarities between the task attributes, we provide a visual analysis of task clustering. Because the task attributes are heterogeneous, the attribute data of tasks are often of multiple types. Taking the early warning tasks as an example, the height is numeric, the target type is ordinal, and the warning type is nominal. To solve this problem, we convert all important attributes to the common interval $[0.0, 0.1]$, and combine these different attributes into the same similarity matrix. If the data set contains P attributes a of mixed types, the similarity distance $d(I, j)$ between tasks i and j is defined as

$$d(i,j) = \frac{\sum_{a=1}^{p} \delta_{ij}^{(a)} d_{ij}^{(a)}}{\sum_{a=1}^{p} \delta_{ij}^{(a)}}, \tag{1}$$

where $\delta_{ij}^{(a)}$ is used to judge the possibility of computing similarity. The indicator $\delta_{ij}^{(a)} = 0$, if either x_{ia} or x_{ja}, is missing (i.e., there is no measurement of attribute a for task i or j); otherwise, $\delta_{ij}^{(a)} = 1$. The contribution of attribute a to the dissimilarity between i and j (i.e., $d_{ij}^{(a)}$), is computed according to its type.

(1) If a is a numeric: $d_{ij}^{(a)} = \frac{|x_{ia} - x_{ja}|}{\max_h x_{ha} - \min_h x_{ha}}$, where h runs over all non-missing objects for attribute a.

(2) If a is nominal or binary: $d_{ij}^{(a)} = 0$ if $x_{ia} = x_{ja}$; otherwise, $d_{ij}^{(a)} = 1$.

(3) If a is ordinal: M_a represents the number of possible states, and r_{ia} represents corresponding rank. Compute ranks r_{ia} and $z_{ia} = \frac{r_{ia} - 1}{M_a - 1}$, and treat z_{ia} as numeric.

Here, the normalization method of different types is given, and the specific similarity measure will be introduced in the active learning algorithm. A similarity matrix is obtained by comprehensive distance

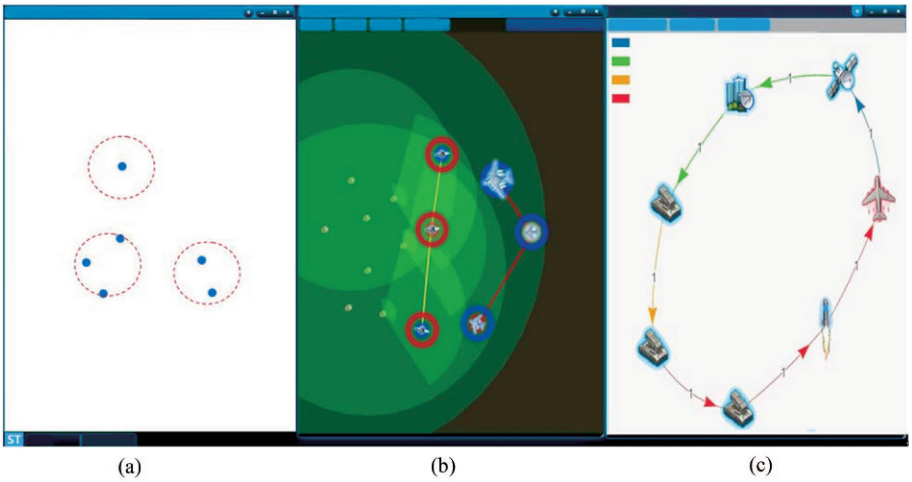

Figure 4 Three-pane interactive interface. (a) Similarity space visualization based on task attribute clustering; (b) visualization of battlefield environment based on task information; (c) visualization of the whole operation effect based on OODA interaction network.

calculation of the three types, and the clustering is then visualized. The clustering of six warning tasks is shown in Figure 4(a). Each point in the space represents a task, and the distance between two points measures the difference between the two tasks. Experts can make an overall judgment on all tasks using this visualization tool and determine the standard for evaluation. Our experiment proves that this method cannot improve both efficiency and accuracy. Moreover, experts can choose a representative task to score by using clustering visualization in their first evaluation, thereby improving learning accuracy and reducing the number of query iterations. However, it is only a measure of the task itself, and cannot reflect the impact of a single task on the entire confrontation. Therefore, we provide a visual interface based on OODA interactive network to reflect the impact of the task on the entire confrontation process and to help the experts produce a comprehensive assessment.

4.2.3 *Whole operation effect visualization based on OODA interaction network*

The OODA loop [41] proposed by the United States Army colonel John Boyd is a kind of high abstraction of the operation process "Observe-Orient-Decide-Act". In accordance with the needs stated in the evaluation of combat missions, we further refine them into 26 kinds of events corresponding to the rules of military affairs, and these events can help us describe the whole process of the operation in detail. Any operational result is produced by the joint actions of these events. By the retrospective analysis of a combat task result, we can arrange the relevant events of the task to form a closed loop, which is essentially the OODA loop. Figure 4(c) shows the OODA loop of one of six warning tasks: "satellite detect aircraft", "satellite transmit information to control center", "control center send command to missile brigade", "missile brigade send command to missile battalion", and "missile battalion fire to aircraft". These events constitute a complete OODA loop. Through this example, we discover that this warning task plays a key role in the final missile attack, and it would be highly evaluated. While some other warning tasks also find aircraft, they do not aid the WSOS in shooting them down for various reasons, and therefore cannot be given a higher evaluation. It is in line with actual operational rules that we evaluate tasks by the final combat efficacy. Our system enables the retrospective analyses of all key events and forms a few OODA loops. We merge the same combat unit involved into each of these loops, and eventually all loops are synthesized into a net, called an OODA interaction network (as shown in Figure 5). This net structure can precisely reflect the actual operational mechanism; that a task may impact a few combat results. Military experts can observe combat efficacy by analyzing the OODA interactive net.

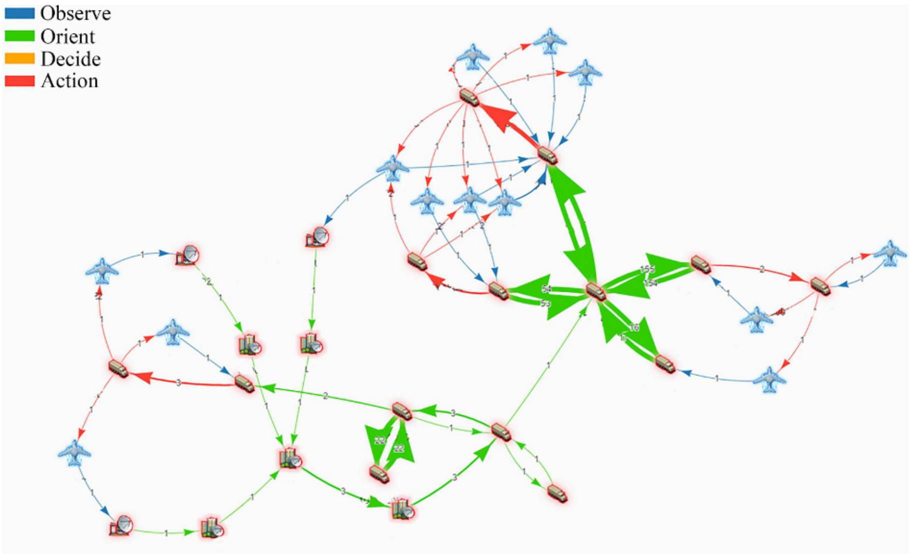

Figure 5 OODA interactive network.

4.3 Active learning algorithm

To learn fast expert evaluation principles with the least number of queries, the system introduces the active learning model to help experts evaluate large-scale tasks. Because our task data comes from multiple simulation results, the system default takes the tasks of the first simulation as the initial selection task set for the experts, and the experts can independently determine the initial number of evaluation tasks; the next evaluation number is provided by the active learning model.

4.3.1 *Similarity measure and optimization calculation*

We use a global similarity matrix to express the relation between tasks for calculating all tasks efficiently in a unified manner. In Subsection 4.2.2, we provide a method for calculating the similarity distance between three different types of tasks. In this section, we provide an additional calculation formula for each element in the similarity matrix:

$$w_{ij} = \exp\left(-\frac{d(i,j)^2}{\sigma^2}\right), \tag{2}$$

σ is a measure parameter of the attribute weight. It can be adjusted in the system, though its default value is 1 in our experiment. $d(i,j)$ is the normalized distance of all attributes, as shown in Subsection 4.2.2. In order to efficiently compute the distance between tasks, we use the framework provided in [42], as this framework can provide the optimal low dimensional embedding to improve computational efficiency; it maps the original distance matrix onto a low dimensional manifold by using the Diffusion Heat technique. We define the global similarity matrix $W = [w_{ij}]$ and G as the global similarity graph. N task instances in all task instances TD $= \{x_1, \ldots, x_n\}$ are represented as points on the G, while the similarity distances w_{ij} between tasks are represented by edges of the G. Per the theorem analysis in [43], the global similarity matrix also shows the transition probability between tasks; therefore, we consider the relation of task scores as the energy conversion on the Gauss field,

$$E_G(y) = \frac{1}{2}\sum_{ij} w_{ij}(y_i - y_j)^2. \tag{3}$$

We define y_i as the value of the scored task i $(i \in l)$ in Subsection 4.1. Now y_i is derived forthe entire task set TD, and it represents the desired value of the task i $(i \in n)$. y is defined as the function of y_i, so the prediction of task evaluation is, in actuality, the process of the solution.

First, according to the nature of the graph, the diagonal matrix of the degree of all nodes on the graph is defined, and the matrix $Z = D^{-1}W$ can then be obtained by the global similarity matrix W.

Second, according to the properties of the harmonic function, the solution of the energy function is $y = \arg\min_{y|L=y_l} E_G(y) = Zy$. Depending on whether the task is scored, W and y can be decomposed into $\begin{bmatrix} W_{ll} & W_{lu} \\ W_{ul} & W_{uu} \end{bmatrix}$ and $\begin{bmatrix} y_l \\ y_u \end{bmatrix}$, respectively. Therefore, we can obtain the solution to the non-evaluation task.

Finally, the algorithm is optimized. Since Z is a real symmetric matrix $D^{-\frac{1}{2}}WD^{-\frac{1}{2}}$, it can be transformed into a matrix, and the normalized eigenvalues are obtained $1 = \lambda_0 \geqslant \lambda_1 \geqslant \cdots > 0$. By taking the first u feature value of the maximum vectors, the similarity matrix is embedded into the low dimensional space for efficient computation.

4.3.2 *Selection strategy*

The key functions of the active learning algorithm in the system are to design sampling criteria for measuring tasks and to select high value tasks for the military personnel to evaluate. The uncertainty of unlabeled samples is always taken as an important selection factor. Entropy is used to measure uncertainty in our system: first, according to the task information evaluated by military personnel, the learning algorithm predicts the unevaluated tasks and obtains the posterior probabilities $p(y_i)$ of each unevaluated task i; second, the information entropy of each unevaluated task is calculated based on the posterior probability.

$$H(i) = -\sum_{y_i \in Y} p(y_i)\log p(y_i). \tag{4}$$

Per the characteristics of information entropy in information theory, greater task information entropy and greater uncertainty of the task evaluation should increase the selection of the task of maximum information entropy for military personnel scoring. However, when facing multiple classification problems (such as the multiple scoring categories of tasks), a few categories of small probability would affect the judgment and accuracy of entropy. Therefore, we use the guidelines BvSB of optimal and suboptimal labeling [44], which consider the most likely two kinds of evaluation results, and ignore other kinds of small probability results. The probabilities of the most and second most possible score of task i are identified as $p(y_i^{\text{Best}})$ and $p(y_i^{\text{Second}})$, and the guidelines can be expressed as follows:

$$\text{BvSB} = \mathop{\arg\min}_{l<i\leqslant u+l}(p(y_i^{\text{Best}}) - p(y_i^{\text{Second}})). \tag{5}$$

In the specific implementation of the algorithm, we convert the multi-classification problem into a few two-classification problems: first, we calculate posterior probability of each score category of tasks separately; then, we solve the different values between $p(y_i^{\text{Best}})$ and $p(y_i^{\text{Second}})$ of tasks; finally, we find the k tasks with the fewest different values in all tasks, and the greatest uncertainty of predictive values where the k value can be adjusted according to the specific tasks of the system; the default is 10.

4.4 Overall presentation of task set evaluation results

By the repeated interaction between the experts and the system, the evaluation of large-scale combat tasks has been realized, and a task data set with evaluation information has been produced. To display the overall information of large scale tasks directly to military personnel, the system provides a visual display of statistical indicators and statistical graphics (as shown in Figure 6).

Based on these displays, military personnel can observe value boundaries, distribution features, central tendency, deviation degrees, and correlation of the evaluation results of the task set, and can then evaluate the overall mission completion of the WSOS. By means of this visual display, military personnel can do further analysis on the evaluation as follows: (i) discover hidden high value operational rules, (ii) analyze and evaluate the reasonableness of scoring, and (iii) comparatively analyze whether the results of active learning are consistent with the scoring principle of estimators.

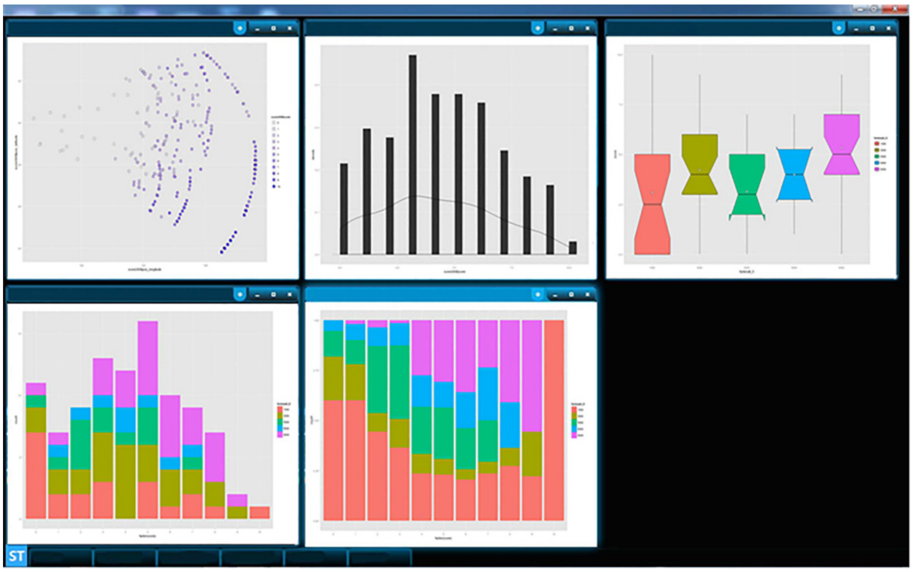

Figure 6 Presentation interface of task set evaluation results.

5 Experiment

The mission evaluation system is mainly applied to the evaluation of large scale combat tasks in the simulation tests of WSOS confrontations. To assess the effect and the role of the actual application in the military, we organized military commanders, academy teachers, and graduated students to carry out the experimental analysis, which considers both the visual interactive and the active learning models. In the task analysis, we use a workstation with eight 3.60 GHz Intel i7-4790 CPUs, 16 GB memory, and an Nvidia GTX 750 graphics card with 4022 MB video memory. We take two combat simulation experiments of different missions. In the reconnaissance mission evaluation experiment, three scale task data sets are used for the first interactive time of active learning: the 100 tasks data set (0.32 s), the 300 tasks data set (0.63 s), the 700 tasks data set (0.88 s). In the fire strike mission evaluation experiment, two kinds of scale tasks data set are used for the first interactive time of active learning: the 100 tasks data set (0.45 s), and the 400 tasks data set (0.75 s).

5.1 Experiment of visual interactive effect

In traditional combat simulation experiments, the military experts evaluate the completion of the task based on analysis of the original battlefield situation information. When using the proposed mission system, the expert can evaluate tasks with the aid of a three-pane visual interface. To accurately measure the effect of visual interaction, we carried out an experimental evaluation of the traditional methods, a partial view of the proposed method, and the complete three-pane method.

5.1.1 *Experiment design*

Data sources. Our experiment simulates the confrontation process of an air defense WSOS warning of the red side against the blue aircraft of the blue side based on the simulation environment of the WSOS. The red warning SOS comprises four fixed radar positions, three mobile radar positions, one radar command node, and one information center, and carries out three kinds of early warning tasks (DET, IFF and NCTR) for the intruding aircraft. Three types of intruding aircraft (A, B and C) fly to the center of the air defense position in an arc formation from 50 different directions (as shown in Figure 7), and their possible flight heights are 1000, 3000, 5000, 6000, and 8000 m. From these simulation experiments, a total of 10800 early warning tasks were produced. We have sampled 20 valid tasks for the experimental analysis.

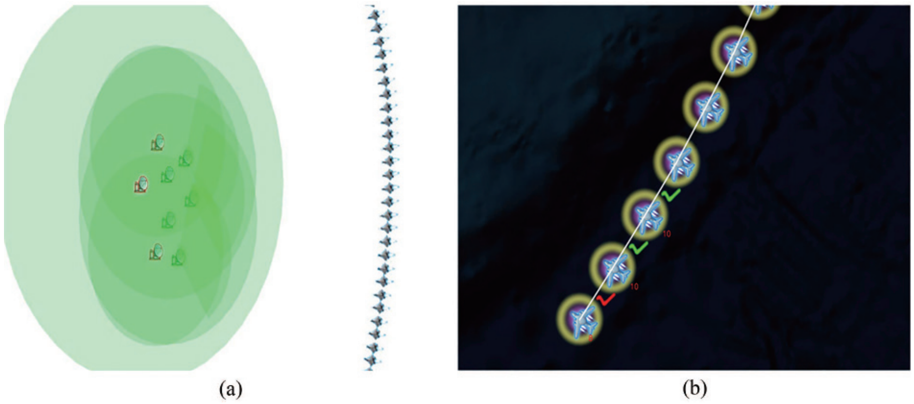

Figure 7 Early warning task experiment for large scale air targets. (a) Overall concept map; (b) military personnel scoring interface.

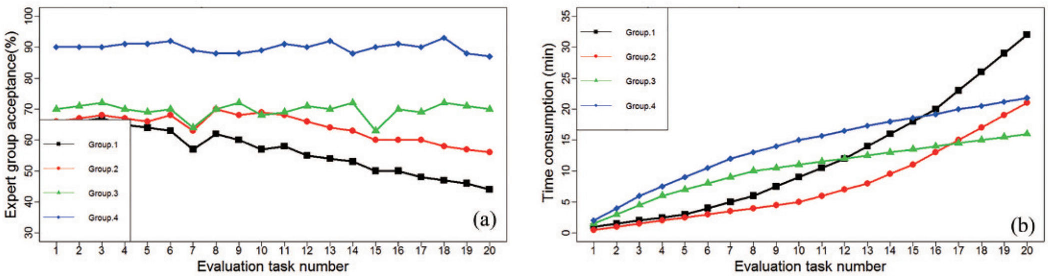

Figure 8 Comparative analysis of different visualization methods. (a) Expert group acceptance of each group evaluation result; (b) time consumption of each group evaluation result.

Test staff. We choose eight military graduate students divided into four groups: the 1st group evaluates relying on the original battlefield information in the traditional way, the 2nd group evaluates with the help of battlefield visualization based on task information, the 3rd group evaluates by using both battlefield visualization and task attribute clustering visualization, and the 4th group uses the complete three-pane visualization method. Each team must complete the evaluation of the 20 warning tasks independently.

Evaluation criteria. We took the recognition and the time efficiency as two important indicators to measure the visual interactive effect. For the measure of recognition, we organized a group of military personnel with different identities, including military commanders, academy teachers, and air defense professional doctors to constitute an air defense expert group.

5.1.2 *Experiment analysis*

Figure 8(a) shows the recognition of the air defense expert group on the task evaluation results of each group. We find that the evaluation recognition of Group 1 is not high, and the evaluation effect falls into a significant downward trend with the increase in the number of evaluation tasks. This is because the increase in tasks disrupts the evaluation criteria of the estimator. In terms of expert recognition, the performance of Group 2 is superior to that of Group 1. The graduate students explained that the battlefield environment visualization based on task information is helpful; however, there is still an inevitable decline of evaluation quality when the number of tasks significantly increases. Group 3 shows a different trend, specifically that evaluation quality does not decline with an increase in the number of tasks. However, the expert group does not recognize the scoring results of individual tasks, mainly due to the overall impact difference on the individual tasks possessing same attributes. In the case of using the complete three-pane interface, the work of Group 4 was highly regarded, and the overall performance was very stable.

Figure 8(b) shows the time that the four groups of graduate students spent on the evaluation of the 20 operational tasks. The time consumption of Group 1 shows that the increase will cause evaluation time to grow exponentially. The more tasks there are, the more difficult it becomes to determine evaluation standards, thus increasing evaluation time. This is another indicator of the significant impact of the number of tasks on traditional evaluation methods. This can be slightly improved, but there is no significant change on the impact of the number of tasks on efficiency. With the aid of similarity space visualization based on task attribute clustering, grasping evaluation standards becomes easier, and the graduate students in Group 3 could make comprehensive judgments for large scale tasks. Although it may take a little more time for evaluating fewer tasks, there was a clear advantage when evaluating many tasks. Group 4 could judge the impact of a task on the overall operation based on the OODA interactive network, thus the time spent on a single task would increase. However, with the aid of clustering visualization, the time efficiency of Group 4 is still better than those of Group 1 or Group 2 when faced with large numbers of tasks. The above experiment shows the quality and efficiency of evaluation was significantly improved when visualized on a three-pane interface, especially in terms of clustering visualization to efficiency and OODA visualization to recognition.

5.2 Experiment of active learning effect

In our previous work, we did not find a case of applied active learning in a combat simulation. However, due to the particularity of military data, there is no open data set to test. To overcome these difficulties, we use the indicator of accuracy to evaluate small scale tasks and the indicator of recognition to evaluate large scale tasks. In addition, the proposed system can aid the expert group in evaluating recognition by using its statistical graphics display. Finally, whether the combat task is likely to be achieved is a very important standard in the military field, so we set the standard of task completion at 6 points.

5.2.1 *Experiment design*

Data source. Based on the simulation results generated by the above experiments, we extracted 100, 300, and 700 effective tasks. To ensure that the data were representative, we used a balanced strategy to evenly distribute all types of attribute values in each task. The expert group evaluated the data set of 100 tasks as the benchmark ahead of time.

Test staff. To improve the credibility of our test, the air defense expert group evaluated tasks directly, and was responsible for the recognition evaluation of the results of active learning. The experts selected evaluation tasks one of three ways: (i) experts randomly selected evaluation tasks (RS); (ii) experts select the representative tasks with the help of the three-pane interface (TVAS); (iii) experts selected the representative tasks with the help of the three-pane interface and the active learning model (TVAQ).

Evaluation criteria. When evaluating the 100 benchmark tasks, we used predictive accuracy (PA) and prediction classification accuracy (PCA) as indicators. When faced with 300 and 700 tasks of non C benchmark results, we used predictive recognition (PR) predictive classification recognition (PCR) as indicators. The indicator values of PA and PCA can be directly obtained by comparison to the machine predictive results and the benchmark results. The recognition indicator values of PR and PCR are the overall qualitative judgments which were made by the expert group based on the statistical features of the task evaluation results of machine learning.

5.2.2 *Experiment analysis*

(1) The experiment analysis of 100 tasks. We selected the 100 tasks with benchmark results as the test data set. Figure 9(a) shows the comparison of the prediction accuracy in three ways when evaluating 100 tasks; random selection produced the worst performance, while the accuracy rate did not necessarily rise with the increase of the task selection number. We can see that the accuracy improvement of task evaluation is significant when the visual aid is employed, and that active learning is the best of the three methods, however the overall learning effect is less prominent when evaluating 100 tasks.

Figure 9(b) shows the trend of predictive classification accuracy using the above three methods. After comparison with prediction accuracy, we found that the classification accuracy of random selection

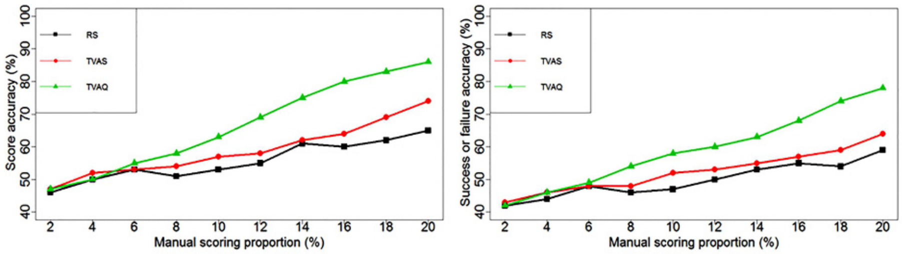

Figure 9 Evaluation experiment of 100 tasks data set. (a) Score accuracy of evaluation based on three machine learning methods; (b) success or failure accuracy of evaluation based on three machine learning methods.

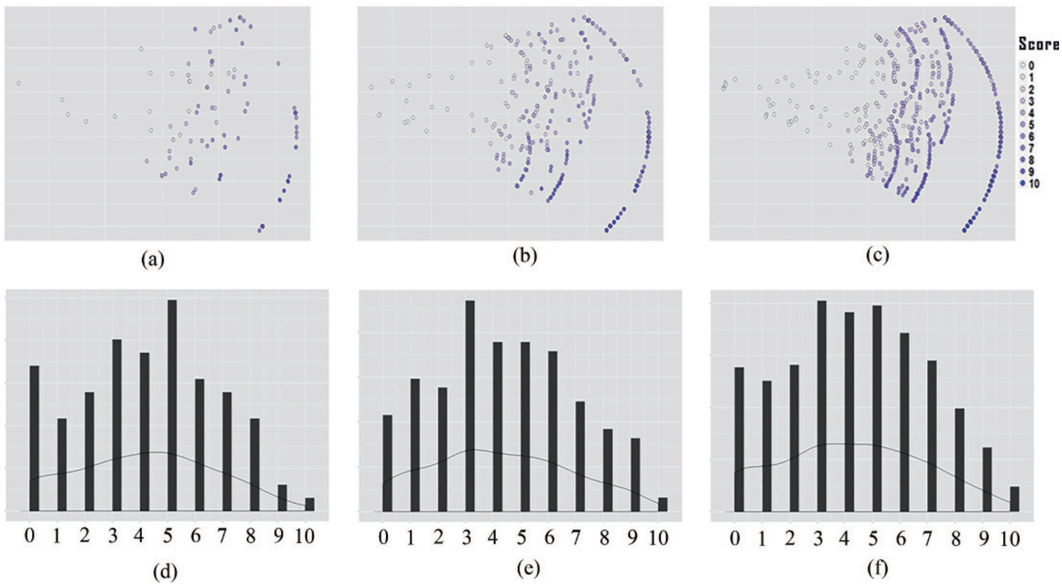

Figure 10 Overall feature analysis. (a) The spatial scatter plot of the 100 tasks evaluation results; (b) the spatial scatter plot of the 300 tasks evaluation results; (c) the spatial scatter plot of the 700 tasks evaluation results; (d) the histogram and density line of the 100 tasks evaluation results; (e) the histogram and density line of the 300 tasks evaluation results; (f) the histogram and density line of the 700 tasks evaluation results.

was relatively static. In addition, the predictive classification accuracy of the visual assistant selection decreased significantly, as the number of tasks in the critical state of classification was not clear; this resulted in a decline in overall predictive classification accuracy. With the active learning method, these tasks in the critical state of classification are again queried by experts, meaning the predictive classification accuracy was better than the predictive accuracy; this shows our objective function played a decisive role.

(2) The recognition analysis of 300 and 700 tasks. The data sets of 300 and 700 tasks have no benchmark evaluation results, so we first analyzed the recognition of the evaluation results of the two task data sets. Our evaluation method was simple and practical; according to the evaluation rules and principles of the 100 tasks data set, we compared the feature of the evaluation results from multiple angles, and the overall recognition of the 300 and 700 task evaluation results was made by the expert group. The emphasis here is that we used a balanced sampling strategy when we extracted all three data sets to ensure the credibility of the proposed method. Next, we demonstrated the evaluation results based on the machine learning mode of a three-pane active query. We first analyzed the overall features of the scoring results of the three task data sets, and then compared the features of each attribute.

Overall feature analysis. we use the three modes of the spatial location scatter plot, density map, and 12 statistical indicators to make a comparative analysis of the overall features of the scoring results.

In the spatial scatter plots shown in Figure 10(a)–(c), each point represents a task, and deeper colors represent higher scores, and vice versa. We found that the evaluation results of the three data sets all

Table 1 12 statistical indicators of evaluation results of three data sets

N	Mean	Var	Std_dev	Median	Std_err	CV	CSS	USS	R	R1	Kurtosis	Skewness
97	4.1	6.5	2.5	4	0.26	61.9	622	2263	10	4	−0.79	0.04
293	4.3	6.6	2.6	4	0.15	60.1	1919	7252	10	4	−0.80	0.17
681	4.3	6.8	2.6	4	0.10	61.0	4630	17073	10	4	−0.83	0.10

showed the same features; the upper left area with low scores, and the lower right with higher scores. This is consistent with the actual situation of our experiment, and the color distribution of the scatter plots of the three data sets is also essentially consistent.

In the histograms and density lines shown in Figure 10(d)–(f), the 11 columns from left to right represent the frequency of the 0–10 score and the curve represents the density values of the corresponding data set. The three data sets show the common feature of high frequency of the middle scores; the difference being that, when compared to the 300 and 700 task data sets, the 0 score of the 100 task set is of higher frequency. Through our analysis, because the distribution of the 100 task set is very sparse, the impact between points becomes weak when propagating in Gauss space, causing many zero value points; while the zero point is not easy to see in the 300 and 700 tasks due to their intensive distribution, which is consistent with the feature of our algorithm.

To produce a more accurate comparison of the overall distribution, we calculated the 12 statistical indicators of the evaluation results of the three task data sets: mean value (Mean), variance (Var), standard deviation (Std_dev), median (Median), standard error (Std_err), coefficient of variation (CV), corrected sum of squares (CSS), UN corrected sum of squares (USS), range (R), semi standard range (R1), coefficient of Kurtosis (Kurtosis), and coefficient of Skewness (Skewness). According to the measurement results, we analyzed the four dominant indicators, namely mean value (Mean), standard deviation (Std_dev), coefficient of Kurtosis (Kurtosis), and coefficient of Skewness (Skewness). There was no significant difference in terms of Mean and Std_dev of the three data sets, which shows that the distribution of each score set was the same. The Kurtosis of the three data sets were in the range [−1,1], so we can use Skewness to measure the distribution difference of evaluation results for the three task sets. The Skewness values of the three data sets are 0.04, 0.17, and 0.10, as shown in Table 1, and we found the features of the 700 tasks evaluation set was closer to the feature of the 100 tasks evaluation set, which was the benchmark result. This shows our algorithm is more accurate with an increasing number of tasks.

Analysis of feature of each task attribute. We compared the evaluation results of the three task sets with different task attribute angles, including the target height attribute (as shown in Figure 11), the target type attribute (as shown in Figure 12), and the task type attribute (as shown in Figure 13). The experimental results showed that the three data sets can essentially reflect the consistent evaluation features of all attribute types. For example, for the target height attribute the overall evaluation result of the 1000 m altitude target was lowest, the overall evaluation result of the 8000 m altitude target was highest, and the overall evaluation results of other high levels were in the middle. For the target type attribute, the overall evaluation result of the type A target was highest. For the task type attribute, the overall evaluation result of DET was the highest, that of IFF was in the middle, and that of NCTR was lowest. These experimental results show that our active learning algorithm can learn well from the rules of expert evaluation.

(3) Comprehensive experimental evaluation to three task data sets. Based on the previous analysis method of the recognition, we conducted a comprehensive evaluation of the learning results of three selection modes, as shown in Figures 14 and 15. First, we took the 300-task data set as an example to observe the change of score predictive recognition and success or failure predictive recognition, with the proportional increase of manual scoring in the same data set. In random selection mode (RS), the score predictive recognition and success or failure predictive recognition were not obviously promoted with the proportional increase of manual scoring. In the three-view assisted selection mode (TVAS), the recognition of score prediction was significantly improved compared those of success or failure prediction. In the three-view active query mode (TVAQ), the system combined the three-pane interface and active learning method to aid expert scoring, and the recognitions of score and success or failure prediction were

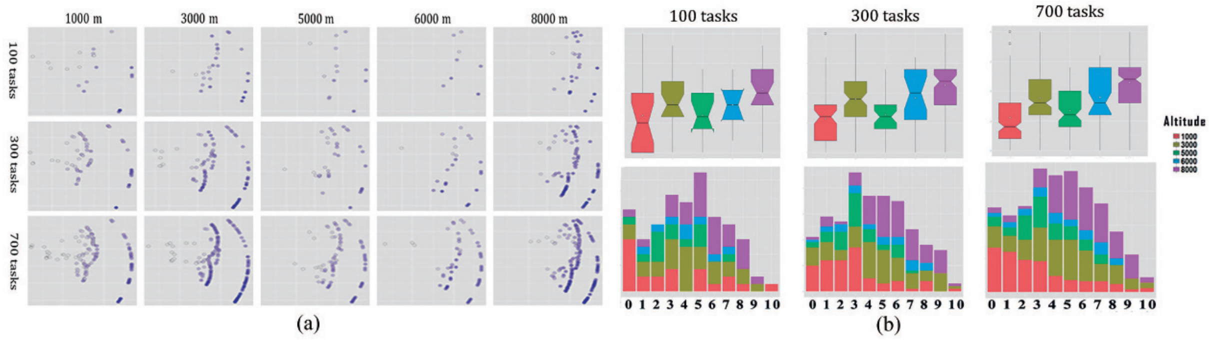

Figure 11 Analysis of different target heights. (a) The spatial scatter plots of the evaluation results of the three data sets; (b) the box line diagram and bar chart of the evaluation results of the three data sets.

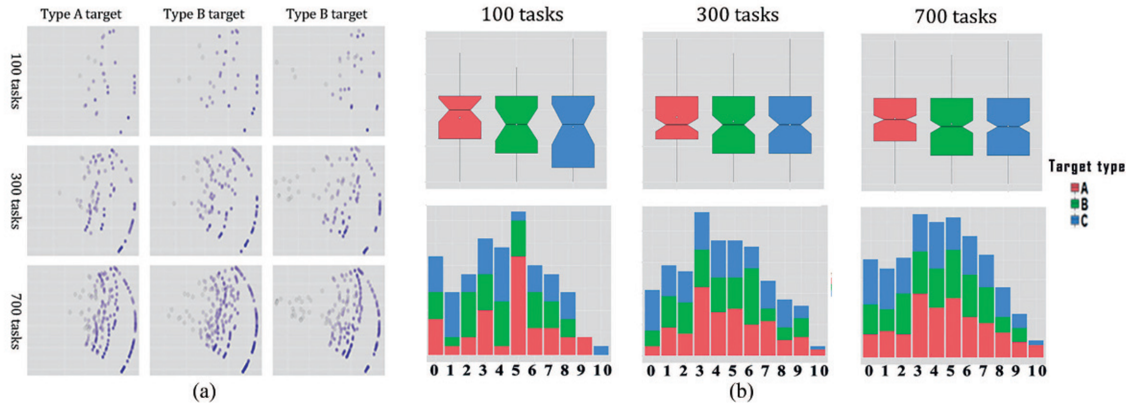

Figure 12 Analysis of different target types. (a) The spatial scatter plots of the evaluation results of the three data sets; (b) the box line diagram and bar chart of the evaluation results of the three data sets.

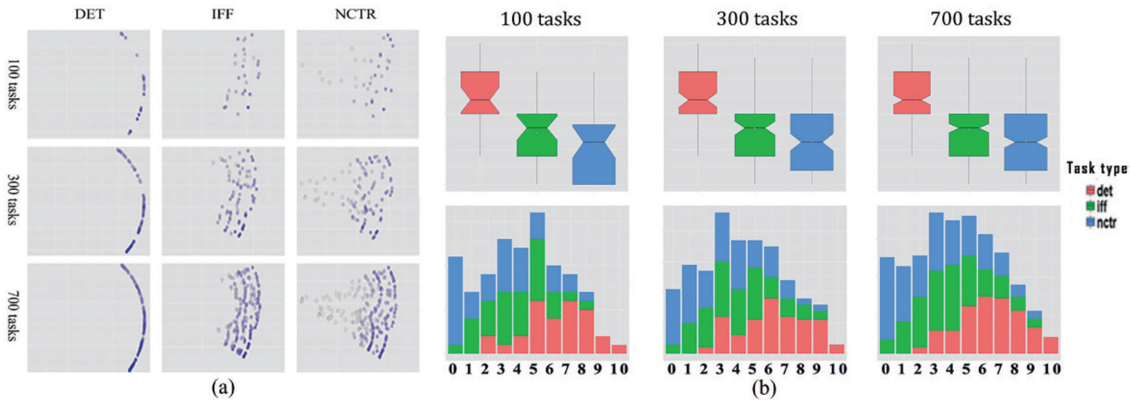

Figure 13 Analysis of the task types. (a) The spatial scatter plots of the evaluation results of the three data sets; (b) the box line diagram and bar chart of the evaluation results of the three data sets.

significantly improved.

Second, we took the 20% manual scoring ratio as an example to observe the performance of the three modes in different task data sets, as shown in Table 2. In random selection mode, the recognition of score prediction of the 100 tasks data set was relatively low, while that of the 300 and 700 tasks sets were almost identical. For each data set, the recognition of success or failure prediction was lower than that of score prediction. In three-view assisted selection mode, the overall performance was better than in the random selection mode, however the overall change trend was same in both modes. In three-view active query mode, the score prediction increases with the number of tasks, and the success or failure prediction recognition was better than the score prediction in the same data set.

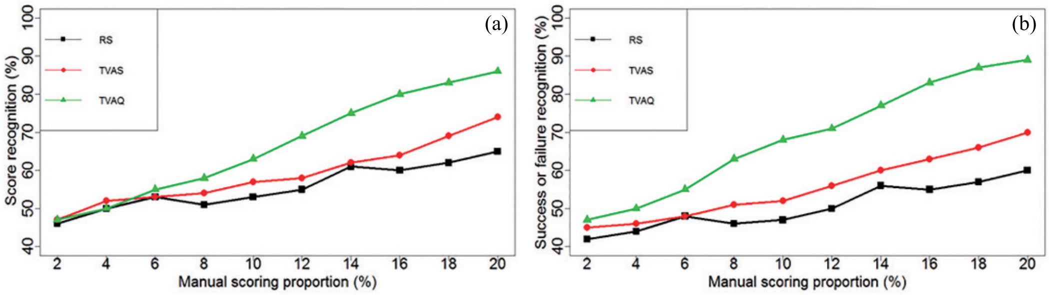

Figure 14 Comparative recognition of different machine learning methods to the 300 tasks data set. (a) The score recognition of evaluation based on three machine learning methods; (b) the success or failure recognition of evaluation based on three machine learning methods.

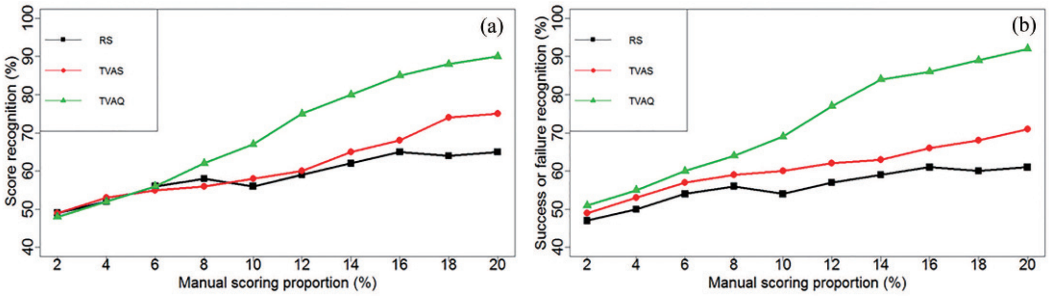

Figure 15 Comparative recognition of different machine learning methods to the 700 tasks data set. (a) The score recognition of evaluation based on three machine learning methods; (b) the success or failure recognition of evaluation based on three machine learning methods.

Table 2 The comprehensive experimental evaluation results of 20% manual scoring

	100 tasks		300 tasks		700 tasks	
	Score recognition	Success or failure recognition	Score recognition	Success or failure recognition	Score recognition	Success or failure recognition
Random selection (RS) three-view (%)	60	59	65	60	65	61
Assisted selection (TVAS) three-view (%)	70	64	74	70	75	71
Active query (TVAQ) (%)	76	78	86	89	90	92

Based on the above comprehensive experimental analysis, we can conclude that the performance of our proposed system's three-view active query mode and success or failure prediction recognition is superior to that of the other examined modes. Therefore, our proposed system is very suitable for the evaluation of large scale tasks.

5.3 The mission evaluation experiment of fire strike

In addition to reconnaissance missions, fire strike is also a common form of military operation. Relying on the mission evaluation system, we evaluated the large-scale combat tasks of the long-range strike WSOS based on a spatial information system, as shown in Figure 16(a). The red strike WSOS includes space satellites, ground stations, control centers and missile forces; the blue side includes some moving targets with anti-missile capability in a certain area. The main attributes of a task include task type, target type, target location, satellite navigation, and anti-missile interference. In our experiments, we sampled groups

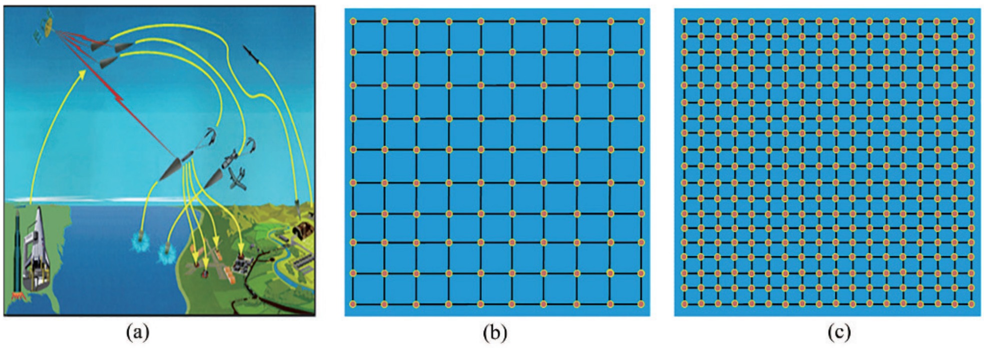

Figure 16 Large-scale and long-range striking tasks experiment. (a) The overall concept map; (b) 100 striking tasks samples; (c) 400 striking tasks samples.

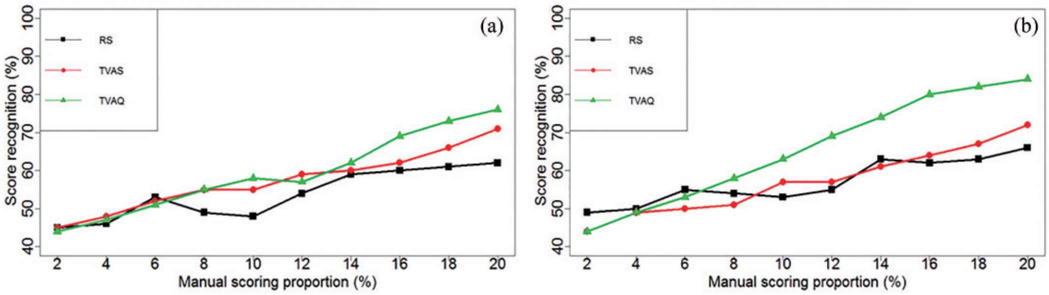

Figure 17 Comparative recognition of different machine learning methods for long-range striking tasks. (a) The score recognition of 100 evaluation tasks based on three machine learning methods; (b) success or failure recognition of 400 evaluation tasks based on three machine learning methods.

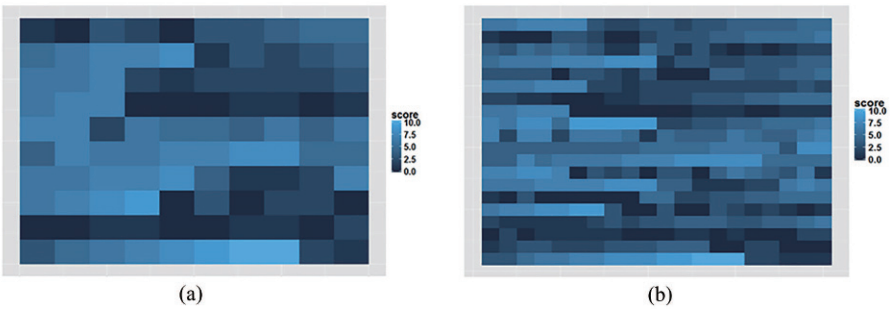

Figure 18 Evaluation results charts of the long-range striking mission based on three-view active query mode. (a) Evaluation results chart of 100 tasks; (b) evaluation results chart of 400 tasks.

of 100 and 400 target points, as shown in Figure 16(b) and (c). Using the above experimental method, we also compared evaluation results based on three machine learning modes to the long-range striking mission, as shown in Figure 17(a) and (b). The experimental results showed that the three-view active query mode of our system also produced the best machine learning effects, and performance improved as the number of tasks increased. Figure 18 shows the evaluation results charts obtained by the three-view active query mode, such as the striking effect on the left side being superior to that on the right, which reveals that the closer the striking distance, the better the missile accuracy.

6 Conclusion

We developed a visual interactive system for mission evaluation of WSOSs to aid military personnel in evaluating large C scale combat tasks. Different from the traditional evaluation methods, the proposed system provides aided visualization of quantitative and qualitative information of tasks based on a three-view interactive interface, a machine learning model for large scale iterative learning based on active inquiry, and visualized presentation for the overall evaluation of missions based on statistical graphics. We demonstrated the function and performance of this mission evaluation system through a large C scale evaluation experiment of two important types of combat missions, proving that the system can not only help military personnel to evaluate large scale tasks efficiently and accurately, but also explain expert evaluation rules and military information intuitively. At present, combat experience is a small data sample; however, operational activities are highly complex. With the help of the improving mission evaluation system, more experience of military commanders and experts can be learned by the machine to assess numerous military tasks and actions. This new evaluation pattern will have a profound impact on future military operation assessment.

Note. Due to confidentiality and space constraints, part of the experimental results and data cannot be displayed in this paper, and some figures were removed from the background pictures, along with some descriptions. Thank you for your understanding.

Acknowledgements This work was supported by Major Program of the National Natural Science Foundation of China (Grant No. U1435218) and National Natural Science Foundation of China (Grant No. 61403401).

Conflict of interest The authors declare that they have no conflict of interest.

References

1 Qian X S, Yu J Y, Dai R W. A new discipline of science—the study of open complex giant system and its methodology. J Syst Eng Electron, 1993, 4: 2–12

2 Dai R W, Li Y D. Researches on hall for workshop of meta synthetic engineering and system complexity. Complex Syst Complex Sci, 2004, 1: 1–24

3 Liu C M, Dai R W. Visualization of experts evaluation opinions in the hall for workshop of meta synthetic engineering. Pattern Recog Artif Intell, 2005, 18: 6–11

4 Dai R W. The proposal and recent development of metasynthetic method(M) from qualitative to quantitative. Chinese J Nature, 2009, 31: 311–314

5 Hu X F, Si G Y. SDE98: an environment of metasynthetic workshop for military strategic decisionmaking. Mini-Micro Syst, 1999, 20: 2–7

6 Si G Y. Strategic decision integration discussion and research and implemenation of the simulation circumstance. Syst Eng, 2000, 18: 79–80

7 Zhang X X, Zhang P Z. Research on visualization of group decision argument opinion's distributing—design and development of electronic common brain audiovisual room. J Manage Sci China, 2005, 8: 15–27

8 Li J, Zhang P Z, Jiang Y Z. Research on automatic topic visual clustering in the group argument support systems. J Syst Manage, 2009, 18: 325–331

9 Xiong C Q, Li D H, Zhang Y. Clustering analysis of experts' opinion and its visualization in hall for workshop of meta-synthetic engineering. Pattern Recogn Artif Intell, 2009, 22: 017

10 Franks A, Miller A, Bornn L, et al. Counterpoints: advanced defensive metrics for NBA basketball. In: Proceedings of the 9th Annual MIT Sloan Sports Analytics Conference, Boston, 2015

11 Maheswaran R, Chang Y H, Henehan A, et al. Deconstructing the rebound with optical tracking data. In: Proceedings of MIT Sloan Sports Analytics Conference, Boston, 2012

12 Lucey P, Bialkowski A, Monfort M, et al. "Quality vs Quantity": improved shot prediction in soccer using strategic features from spatiotemporal data. In: Proceedings of the 8th Annual MIT Sloan Sports Analytics Conference, Boston, 2014. 1–9

13 Rusu A, Stoica D, Burns E. Analyzing soccer goalkeeper performance using a metaphor-based visualization. In: Proceedings of the 15th International Conference on Information Visualization. Washington: IEEE Computer Society, 2011. 194–199

14 Rusu A, Stoica D, Burns E, et al. Dynamic visualizations for soccer statistical analysis. In: Proceedings of the 14th International Conference on Information Visualization. Washington: IEEE Computer Society, 2010. 207–212

15 Pileggi H, Stolper C D, Boyle J M, et al. Snapshot: visualization to propel ice hockey analytics. IEEE Trans Visual Comput Graph, 2012, 18: 2819–2828

16 Perin C, Vuillemot R, Fekete J D. Soccer Stories: a kick-off for visual soccer analysis. IEEE Trans Visual Comput Graph, 2013, 19: 2506–2515

17 Janetzko H, Sacha D, Stein M, et al. Feature-driven visual analytics of soccer data. In: Proceedings of IEEE Conference on Visual Analytics Science and Technology. Washington: IEEE Computer Society, 2014. 13–22

18 Angluin D. Queries and concept learning. Mach Learn, 1988, 2: 319–342

19 Tomanek K, Olsson F. A web survey on the use of active learning to support annotation of text data. In: Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing, Boulder, 2009. 45–48

20 Settles B. Active Learning Literature Survey. Computer Sciences Technical Report 1648, University of Wisconsin-Madison, 2009
21 Dasgupta S, Langford J. A tutorial on active learning. In: Proceedings of International Conference of Machine Learning, Quebec, 2009
22 Macskassy S A. Using graph-based metrics with empirical risk minimization to speedup active learning on networked data. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, 2009. 597–606
23 Bilgic M, Mihalkova L, Getoor L. Active learning for networked data. In: Proceedings of the 27th International Conference on Machine Learning, Haifa, 2010. 79–86
24 Shi L, Zhao Y, Tang J. Batch mode active learning for networked data. ACM Trans Intell Syst Tech, 2012, 3: 33
25 Hu X, Tang J, Gao H, et al. ActNet: active learning for networked teats inmicro-blogging. In: Proceedings of the 2013 SIAM International Conference on Data Mining, Austin, 2013. 306–314
26 Cesa-Bianchi N, Gentile C, Vitale F, et al. Active learning on trees and graphs. arXiv:1301.5112, 2013
27 Chen Y, Li Z, Nie L, et al. A semi-supervised network model for micro blog topic classification. In: Proceedings of the 24th International Conference on Computational Linguistics, Mumbai, 2012. 561–576
28 Liu F C. Research on entity relation extraction technology based on text. National Defense Science and Technology University, 2013
29 Tong S, Chang E. Support vector machine active learning for image retrieval. In: Proceedings of the 9th ACM International Conference on Multimedia, Ottawa, 2001. 107–118
30 Hu X, Tang L, Tang J, et al. Exploiting social relations for sentiment analysis in micro blogging. In: Proceedings of the 6th ACM International Conference on Web Search and Data Mining, Rome, 2013. 537–546
31 Yan R, Yang J, Hauptmann A. Automatically labeling video data using multi-class active learning. In: Proceedings of 9th IEEE International Conference on Computer Vision, Nice, 2003. 516–523
32 Zhu X, Zhang P, Lin X, et al. Active learning from stream data using optimal weight classifier ensemble. IEEE Trans Syst Man Cybernet, 2010, 40: 1607–1621
33 Zhang C, Chen T. An active learning framework for content-based information. IEEE Trans Multimed, 2002, 4: 260–268
34 PraBni J S, Ropinski T, Hinrichs K. Uncertainty—aware guided volume segmentation. IEEE Trans Vis Comput Graph, 2010, 16: 1358–1365
35 Top A, Hamarneh G, Abugharbieh R. Active Learning for Interactive 3D Image Segmentation Berlin: Springer, 2011. 603–610
36 Gao L, Cao Y P, Lai Y K, et al. Active exploration of large 3d model repositories. IEEE Trans Visual Comput Graph, 2015, 21: 1390–1402
37 Cohen-Or D, Zhang H. From inspired modeling to creative modeling. Visual Comput, 2016, 1: 7–14
38 Hofmann T, Puzicha J. Latent class models for collaborative filtering. In: Proceedings of the 16th International Joint Conference on Artificial Intelligence. San Francisco: Morgan Kaufmann Publishers Inc, 1999. 99: 688–693
39 Jin R, Si L. A Bayesian approach toward active learning for collaborative filtering. In: Proceedings of the 20th Conference on Uncertainty in Aartificial Intelligence, Banff, 2004. 278–285
40 Harple A S, Yang Y. Personalized active learning for collaborative filtering. In: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Singapore, 2008. 91–98
41 Boyd J R. A discourse on winning and losing. Maxwell Air Force Base: Air University Library, Document No. M-U 43947 (Briefing slides)
42 Nadler B, Lafon S, Coifman R R, et al. Diffusion maps, spectral clustering and eigen functions of Fokker-Planck operators. In: Proceedings of Neural Information Processing Systems, British Columbia, 2005. 955–962
43 Zhu X J, Lafferty J, Ghahramani Z. Combining active learning and semi-supervised learning using Gaussian fields and harmonic functions. In: Proceedings of International Conference on Machine Learning Workshop, Washington, 2003. 58–65
44 Joshi A J, Porikli F, Papanikolopoulos N. Multi-class active learning for image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Miami, 2009. 2372–2379