Information Sciences 436–437 (2018) 268–281

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

An adaptive decision-making method with fuzzy Bayesian reinforcement learning for robot soccer

Haobin Shi[a,*], Zhiqiang Lin[a], Shuge Zhang[a], Xuesi Li[a], Kao-Shing Hwang[b]

[a] *School of Computer Science, Northwestern Polytechnical University, Xi'an, Shaanxi Province 710072, China*
[b] *Department of Electrical Engineering, National Sun Yat-sen University, Kaohsiung 80424, Taiwan*

ARTICLE INFO

Article history:
Received 24 June 2016
Revised 10 January 2018
Accepted 14 January 2018
Available online 17 January 2018

Keywords:
Situation evaluation
Robot soccer
Fuzzy Bayesian
Reinforcement learning

ABSTRACT

A robot soccer system is a typical complex time-sequence decision-making system. Problems of uncertain knowledge representation and complex models always exist in robot soccer games. To achieve an adaptive decision-making mechanism, a method with fuzzy Bayesian reinforcement learning (RL) is proposed in this paper. To extract the features utilized in the proposed learning method, a fuzzy comprehensive evaluation method (FCEM) is developed. This method classifies the situations in robot soccer games into a set of features. With the fuzzy analytical hierarchy process (FAHP), the FCEM can calculate the weights according to defined factors for these features, which comprise the dimensionality of the state space. The weight imposed on each feature determines the range of each dimension. Through a Bayesian network, the comprehensively evaluated features are transformed into decision bases. An RL method for strategy selection over time is implemented. The fuzzy mechanism can skillfully adapt experiences to the learning system and provide flexibility in state aggregation, thus improving learning efficiency. The experimental results demonstrate that the proposed method has better knowledge representation and strategy selection than other competing methods.

1. Introduction

Robot soccer game is an example of a multi-agent cooperative confrontation platform where multiple agents need to complete complex tasks within a dynamic and uncertain environment [11,2]. A complete decision-making system for robot soccer games comprises two parts: situation evaluation and decision-making [37]. Situation evaluation makes assessment according to the uncertain environmental information and multi-robot situations. Decision-making involves selecting the most appropriate strategy on the basis of the assessment made of the situation in order to achieve a favorable outcome. These two parts are the key technologies of a robot soccer system and have attracted much research interest.

Multi-attribute expert decision-making methods have been frequently used for decision-making [37,32]. The experience knowledge can be effectively expressed by a multi-attribute expert decision-maker. However, the decision-maker is susceptible to a lack of experience and may break down when the available experience does not cover well the experience domain. Hence, this method may perform poorly in dynamic scene evaluation such as robot soccer games [15]. To resolve this problem, another method called the knowledge representation and rule-based reasoning method with fuzzy logic is applied to dynamic decision-making [33,25,39]. This approach can effectively quantify the expert's experience and make the agent's

[*] Corresponding author.
E-mail address: shihaobin@nwpu.edu.cn (H. Shi).

https://doi.org/10.1016/j.ins.2018.01.032

decision-making process more in line with human thinking. However, many effective training samples are necessary and the results may be difficult to quantify for decision inference. The use of only a few samples may lead to a slightly lower efficiency, hence, this method may be not appropriate for evaluating a dynamic scene. Knowledge inference methods, such as the blackboard model, logic template matching and Bayesian network inference technology, are employed for decision-making in some studies [19,28]. Knowledge inferred using a Bayesian network has the ability to deal objectively with the uncertainty problem [19,3,9]. It can accurately explain the causal relationship and degree of correlation between the variables, contributing to predict the occurrence of events through directed edges and conditional probability distribution. The information of each node for the Bayesian network has a significant impact on the inference results. Since its representation ability for the input data is weak, node information has a strong influence on the results of inference. Thus when the input information is insufficient for quantitative analysis, this method may be less appropriate and it is limited in its application to the evaluation of robot soccer under such a complex dynamic environment.

For strategy selection, a Bayesian network is usually applied [19,3]. It is sensitive to *a priori* knowledge, which is normally difficult to be acquired, and hence cannot work well in a dynamic environment. Therefore, the single Bayesian network-based strategy selection method is somewhat limited in its applications. A neural network is an alternative for conducting strategy selection [12]. Neural networks have good abilities of linear and nonlinear mappings and self-adaptation, so that they can be applied to strategy selection. However, the method of neural network-based strategy selection needs a large amount of offline training and has high complexity, making it a shortcoming for dynamic strategy selection. Reinforcement learning (RL) systems have been applied to selecting the most appropriate decision scheme in some applications [7]. After a period of learning in different environments, this method can achieve adaptive decision-making with a high degree of adaptability, robustness, and versatility. Another effective method for decision-making of autonomous robots is batch reinforcement learning, which can be adapted to varying requirements in a variety of scenarios [26]. In addition, a self-learning cooperative strategy for robot soccer systems has been developed using a combination of adaptive q-learning and fuzzy method. The objective of the fuzzy method in this work is to evaluate states and rewards, and all the learning parts are dependent on adaptive q-learning [8]. A method developed from Q value-based dynamic programming with multi-agent reinforcement learning is used for route planning by combining Q-value with Boltzmann distribution to create a priority route plan [40].

In summary, the decision making process can be challenging when the environment is complicated due to a lack of necessary information required for decision-making. Traditional logic requires additional axioms and constraints to deal with the real world as opposed to the ideal world of mathematics. Certainty factors are associated with rules and conclusions, which is referred to as fuzzy logic. Furthermore, current methods have limitations in strategy selection that reduces their environmental adaptability. This study proposes an adaptive decision-making method that involves fuzzy Bayesian reinforcement learning and applies it to robot soccer games. Prior to utilizing the proposed method, which is an extension from FCEM, a fuzzy situation evaluation method is developed for assessing competition situations. In this approach, situations are divided into two classes: real-time and sequential situations. By means of FAHP [10,23], weights of evaluation factor imposed on situations can be obtained. With the comprehensively evaluated situations, a Bayesian network for situation forecast is applied to calculating decision bases. For adaptive decision-making, an RL learning system is designed by regarding the outputs of the Bayesian network as states. This paper is organized into six sections. Following the Introduction section, the overall framework of the proposed method is shown in Section II. Section III presents the fuzzy situation evaluation method where the separated situations and FAHP are imported into the FCEM. Section IV puts forward an adaptive decision-making method that involves Bayesian-RL where states are calculated using Bayesian network and strategies are selected through RL. To illustrate the performance of the proposed method, an experimental comparison is given in Section V. Section VI presents the conclusions.

2. Framework of proposed adaptive decision-making method

In a complex sequential decision-making system, a reasonable situation evaluation and an adaptive decision-maker are important for a robot soccer system. The framework of the proposed method is shown in Fig. 1. In the specific time segment *T*, the environmental changes are recorded according to the correlation of the observed feature values, while environmental data are filtered and analyzed automatically to provide a reasonable explanation for the current situation. According to the analysis of situation, the adaptive decision-making subsystem can choose an appropriate strategy. This study divides the decision-making system into four steps: situation awareness, situation understanding, situation forecast, and adaptive decision-making [36], as shown in Fig. 1.

Situational awareness module. This is the first layer of the framework. According to the multi-factor evaluation method and relevant domain knowledge, this proposed method extracts situation evaluation factors, namely remaining time, difference in scores, competition-controlling ability, ball-controlling ability, shooting ability, ball position, and the difference in distance to the ball between us and the opponent (DDB).

Situational understanding module. In this module, the forecast information is quantized through a robot soccer class. FAHP [6] is employed to establish a weight distribution library and construct a knowledge-based situation understanding module by combining training data with domain expert knowledge.

Situation forecast module. Inference of the decision bases is made using a fuzzy Bayesian network that takes ball position, DDB and situation understanding knowledge as the main evaluation factors.

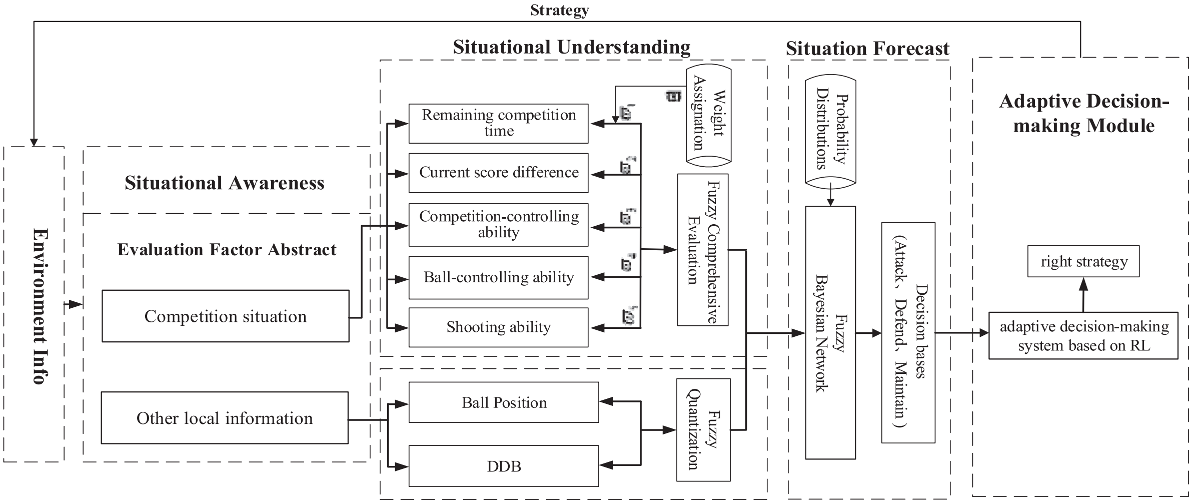

Strategy
Situational Understanding
Situation Forecast
Adaptive Decision-making Module
Environment Info
Situational Awareness
Evaluation Factor Abstract
Competition situation
Other local information
Remaining competition time
Current score difference
Competition-controlling ability
Ball-controlling ability
Shooting ability
Weight Assignation
Fuzzy Comprehensive Evaluation
Ball Position
DDB
Fuzzy Quantization
Probability Distributions
Fuzzy Bayesian Network
Decision bases (Attack、Defend、Maintain)
right strategy
adaptive decision-making system based on RL

Fig. 1. The framework of the proposed method.

Adaptive decision-making module. According to the calculated decision bases determined using the Bayesian network, the adaptive decision-making system that involves RL is implemented to obtain the most appropriate strategy.

3. Fuzzy situation evaluation method

Since situation evaluation and situation awareness are the basis for intelligent decision-making [20,4], this paper presents a situation evaluation architecture, where evaluated situations are categorized into real-time and sequential situations. Furthermore, as an expansion of the FCEM [34], this study designs a fuzzy situation evaluation method for assessing the situation. The situation understanding information obtained comprises the selection of factors, the number of factors and the type of factors that had to be considered. There is trade-off problem in determining the number of factors. If there are only a few factors, the situation evaluation model may be incomplete in describing the environment. Contrarily, the situation evaluation model with too many factors may have poor real-time performance. Moreover, the type of factors has a significant impact on situation evaluation. With domain expert knowledge, the five most influential factors that affect the competitive situation in robot soccer are selected [29]. The factors are remaining competition time, current score difference, competition-controlling ability, ball-controlling ability and shooting ability.

3.1. Real-time evaluation factors

Assuming that current time is the t^{th} time slice in the robot soccer platform, the proposed method uses the current remaining competition time η^t and the current score difference S^t as the real-time evaluation factors.

Definition 1. Remaining competition time.

$$\eta^t = T_{\max} - t \tag{1}$$

where $T_{\max}$ denotes the number of longest time slices of one competition.

Definition 2. Current score difference.

$$S^t = S^t_m - S^t_o \tag{2}$$

where S^t_m denotes my current score and S^t_o denotes the opponent's current score.

3.2. Sequential evaluation factors

Assuming that the current time is the t^{th} time slice, this study defines the competition-controlling ability φ^t, ball-controlling ability c^t, and shooting ability u^t as the sequential evaluation factors that are calculated using the data in a period of time.

Definition 3. Competition-controlling ability.

$$\varphi^t = b_m/(b_m + b_o + \varepsilon_\varphi) \tag{3}$$

where b_m represents the number of times the ball is in our court from the $(t-T)^{th}$ time slice to the t^{th} time slice; b_o represents the number of times when the ball is in the opponent's court; ε_φ is a constant for ensuring that he denominator

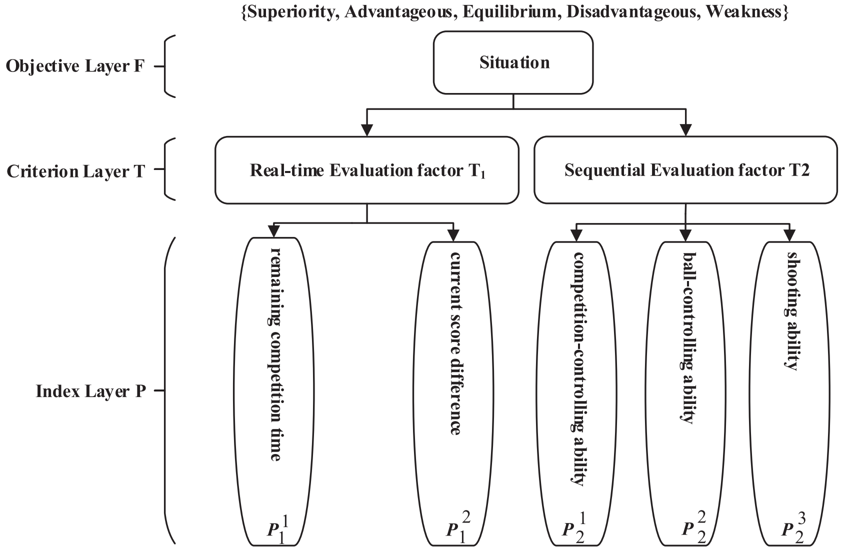

{Superiority, Advantageous, Equilibrium, Disadvantageous, Weakness}
Objective Layer F
Situation
Criterion Layer T
Real-time Evaluation factor T1
Sequential Evaluation factor T2
Index Layer P
remaining competition time
P_1^1
current score difference
P_1^2
competition-controlling ability
P_2^1
ball-controlling ability
P_2^2
shooting ability
P_2^3

Fig. 2. Diagram of the designed hierarchy.

is not zero, and T is also a constant for limiting the length of time. The values of ε_φ and T should be determined by the actual environment.

Definition 4. Ball-controlling ability.

To calculate the ball-controlling ability, the judgment method of ball-controlling for the t^{th} robot should be designed. In the t^{th} time slice, with the robot as the center, whether the t^{th} robot is controlling the ball or not can be determined by

$$f_i^t(x_b^t, y_b^t, x_i^t, y_i^t) = \begin{cases} 1, \sqrt{(x_b^t - x_i^t)^2 + (y_b^t - y_i^t)^2} \le r_c \\ 0, \sqrt{(x_b^t - x_i^t)^2 + (y_b^t - y_i^t)^2} > r_c \end{cases} \tag{4}$$

where r_c is a constant that should be determined by the actual environment; (x_b^t, y_b^t) is the global coordinate of the ball, and (x_i^t, y_i^t) is the global coordinate of the t^{th} robot. The value 1 of $f_i^t(x_b^t, y_b^t, x_i^t, y_i^t)$ indicates that the t^{th} robot is controlling the ball. Or conversely, that the t^{th} robot is not controlling the ball.

Therefore, from the $(t-T)^{th}$ time slice to the t^{th} time slice, the ball-controlling ability is defined as

$$c^t = c_m/(c_m + c_o + \varepsilon_c) \tag{5}$$

where c_m represents the number of times that the ball is controlled by us, i.e. $c_m = \sum_{j=0}^{T}\sum_{i=1}^{n_r} f_i^{t-j}(x_b^t, y_b^t, {}^m x_i^t, {}^m y_i^t)$ where $({}^m x_i^t, {}^m y_i^t)$ is the coordinate of our t^{th} robot; c_o represents the number of times that the ball is controlled by the opponent, i.e. $c_o = \sum_{j=0}^{T}\sum_{i=1}^{n_r} f_i^{t-j}(x_b^t, y_b^t, {}^o x_i^t, {}^o y_i^t)$ where $({}^o x_i^t, {}^o y_i^t)$ is the coordinate of opponent's t^{th} robot; n_r is the number of robots; and ε_c is a constant that should be determined by the actual environment to ensure the denominator is not zero.

Definition 5. Shooting ability.

$$u^t = u_m/(u_m + u_o + \varepsilon_u) \tag{6}$$

where u_m represents the number of our shootings from the $(t-T)^{th}$ time slice to the t^{th} time slice; u_o represents the number of the opponent's shootings, and $\varepsilon_u$is a constant for ensuring that the denominator is not zero.

In particular, when $t < T$, the evaluation factors should be calculated from the 0^{th} time slice to the t^{th} time slice.

3.3. Fuzzy comprehensive situation evaluation algorithm (FCSEA)

The weights for evaluation factors affect the performance of a fuzzy comprehensive evaluation; hence, the FCSEA, in contrast to the traditional experience-based assignment methods [37,1], utilizes an FAHP [4,23] method that estimates the weights by dividing the evaluation factors into different layers.

As shown in Fig. 2, the hierarchical structure, which comprises three layers: index layer, criterion layer, and objective layer, is designed using the decomposition method [30,5]. The objective layer is the ultimate goal i.e. the competition situation, which is expressed as five descriptions: {Superiority, Advantageous, Equilibrium, Disadvantageous, and Weakness}. The criterion layer determines the types of criteria for the target, namely real-time and sequential situations. The index layer shows the five evaluation factors with the constraint of the objective layer and the criterion layer [6].

When conducting fuzzy comprehensive evaluation, the precedence relation matrixes should first be constructed. According to Fig. 2, the relative importance of every two nodes to their common parent node can be obtained through numerous experiments. Hence, three fuzzy importance matrixes which are complementary matrixes are constructed as $R^T = [r_{ij}^T]_{2\times2}$, $R^{P_1} = [r_{ij}^{P_1}]_{2\times2}$, and $R^{P_2} = [r_{ij}^{P_2}]_{3\times3}$. Then these fuzzy importance matrixes are transformed into fuzzy consistent matrixes $R^{T'} = [r_{ij}^{T'}]_{2\times2}$, $R^{P_1{}'} = [r_{ij}^{P_1{}'}]_{2\times2}$, and $R^{P_2{}'} = [r_{ij}^{P_2{}'}]_{3\times3}$ by

$$\begin{cases} r_i = \sum_{k=1}^{n} r_{ik} \\ r'_{ij} = 0.5 + (r_i - r_j)/2n \end{cases}, \quad 1 \le i \le n \tag{7}$$

where n is the rank of one importance matrix; and r_{ik} and r'_{ij} denote the element of one importance matrix or fuzzy consistent matrix, respectively.

According to the fuzzy consistent matrix, the weight of one node A relative to its parent node B can be calculated as follows:

$$\varpi_A^B = 1/n - 1/(2\kappa) + \sum_{j=1}^{n} r'_{i_A j}/(n\kappa) \tag{8}$$

where $k \ge (n-1)/2$ is a constant and should be determined by the actual environment, $\{r'_{i_A j}, j = 1, ..., n\}$ are the i_A^{th} row elements relative to node A in the fuzzy consistent matrix. For instance, $\varpi_{P_1^1}^{T_1} = 1/2 - 1/(2\alpha) + \sum_{j=1}^{2} r_{1j}^{P_1{}'}/(2\alpha)$; $r_{1j}^{P_1{}'} = 0.5 + (r_1^{P_1} - r_j^{P_1})/(2 \times 2)$; and $r_i^{P_1} = \sum_{k=1}^{2} r_{ik}^{P_1}$.

From the designed hierarchy, the weights of the leaf nodes relative to the root node, i.e. the comprehensive weights, can be obtained by a link multiplication where

$$\boldsymbol{\varpi} = (\varpi_{P_1^1}^F, \varpi_{P_1^2}^F, \varpi_{P_2^1}^F, \varpi_{P_2^2}^F, \varpi_{P_2^3}^F) = (\varpi_{P_1^1}^{T_1} \cdot \varpi_{T_1}^F, \varpi_{P_1^2}^{T_1} \cdot \varpi_{T_1}^F, \varpi_{P_2^1}^{T_2} \cdot \varpi_{T_2}^F, \varpi_{P_2^2}^{T_2} \cdot \varpi_{T_2}^F, \varpi_{P_2^3}^{T_2} \cdot \varpi_{T_2}^F) \tag{9}$$

Then the weight vector is regarded as $\boldsymbol{\varpi} = [\varpi_i]_{1\times5} = (\varpi_{P_1^1}^F, \varpi_{P_1^2}^F, \varpi_{P_2^1}^F, \varpi_{P_2^2}^F, \varpi_{P_2^3}^F)$; the evaluation factor vector is defined as $\mathbf{X}^t = [x_i^t]_{5\times1} = (\eta^t, S^t, \varphi^t, c^t, u^t)^T$; and the situation vector is expressed as $Y = [y_j]_{1\times5} =$ (Superiority, Advantageous, Equilibrium, Disadvantageous, Weakness). With inspiration from the fuzzy comprehensive evaluation theory, the membership degree of evaluation factor x_i^t relative to situation y_i can be obtained through a large number of experiments with the fuzzy statistics method. The fuzzy mapping is designed as

$$f : \mathbf{X}^t \to \mathbf{Y}, x_i^t | \to f(x_i^t) \Leftrightarrow (\beta_{i1}^t, \beta_{i2}^t, \beta_{i3}^t, \beta_{i4}^t, \beta_{i5}^t) \tag{10}$$

where $f(x_i^t)$ is the fuzzy evaluation vector of the factor x_i^t relative to the competition situations and β_{ij}^t represents the membership degree of x_i^t relative to y_i.

Therefore, the fuzzy relation matrix in the t^{th} time slice can be obtained by the mapping relationship f that

$$\boldsymbol{\beta}^t = \begin{bmatrix} f(x_1^t) \\ f(x_2^t) \\ \vdots \\ f(x_5^t) \end{bmatrix} = \begin{bmatrix} \beta_{11}^t & \beta_{12}^t & \cdots & \beta_{15}^t \\ \beta_{21} & \beta_{22}^t & \cdots & \beta_{25}^t \\ \vdots & \vdots & & \vdots \\ \beta_{51}^t & \beta_{52}^t & \cdots & \beta_{55}^t \end{bmatrix} \tag{11}$$

The final fuzzy evaluation vector $\mathbf{B}^t = [b_i^t]_{1\times5}$ can be calculated through a "max-min compose operation" with

$$\begin{cases} \mathbf{B}^t = \boldsymbol{\varpi} \cdot \boldsymbol{\beta}^t \\ b_i^t = \vee_{j=1}^{5} (\varpi_j \wedge \beta_{ji}^t) \end{cases} \tag{12}$$

where $\vee$ is the operation for fining the maximum value; and $\wedge$ is the operation for fining the minimum value.

According to the maximum membership degree method, the final comprehensive evaluation result, i.e., the current competition situation, is the description $y_{i_{\max}}$, where $b_{i_{\max}} = \max\{b_i, i = 1, ..., i_{\max}, ..., 5\}$. The corresponding FCSEA is shown in Algorithm 1.

4. Adaptive decision-making method involving Bayesian-RL

Fuzzy logic is better than a Bayesian network for knowledge representation from the viewpoint of human thinking, whereas a Bayesian network is superior to fuzzy logic in the ability of inference and can be adaptive by updating the node probability. To alleviate the inference of uncertain knowledge on the system, this study combines these two methods to perform the situation forecast and output the predicted results, i.e., the decision bases to be incorporated into the RL model for learning.

Algorithm 1 FCSEA.

```
Definition
$\mathbf{B}^t$ := fuzzy evaluation vector in $t^{\text{th}}$ time slice
$i_{\max}$ : = index of maximum value of one vector
$\text{CS}^t$ := current situation in $t^{\text{th}}$ time slice
find_max_index( ): = calculating the index of maximum value
Initialization
    $r_{ij}^{T'} \leftarrow 0.5 + (\sum_{k=1}^{2} r_{ik}^{T} - \sum_{k=1}^{2} r_{jk}^{T})/(2 \times 2)$, for $1 \le i, j \le 2$;
    $r_{ij}^{P_1\,'} \leftarrow 0.5 + (\sum_{k=1}^{2} r_{ik}^{P_1} - \sum_{k=1}^{2} r_{jk}^{P_1})/(2 \times 2)$, for $1 \le i, j \le 2$;
    $r_{ij}^{P_2\,'} \leftarrow 0.5 + (\sum_{k=1}^{3} r_{ik}^{P_2} - \sum_{k=1}^{3} r_{jk}^{P_2})/(2 \times 3)$, for $1 \le i, j \le 3$;
      $\varpi_{T_1}^{F} \leftarrow 1/2 - 1/(2\alpha) + \sum_{j=1}^{2} r_{1j}^{T\,'}/(2\alpha)$;
      $\varpi_{T_2}^{F} \leftarrow 1/2 - 1/(2\alpha) + \sum_{j=1}^{2} r_{2j}^{T\,'}/(2\alpha)$;
      $\varpi_{P_1^1}^{T_1} \leftarrow 1/2 - 1/(2\alpha) + \sum_{j=1}^{2} r_{1j}^{P_1\,'}/(2\alpha)$;
      $\varpi_{P_1^2}^{T_1} \leftarrow 1/2 - 1/(2\alpha) + \sum_{j=1}^{2} r_{2j}^{P_1\,'}/(2\alpha)$;
      $\varpi_{P_2^1}^{T_1} \leftarrow 1/3 - 1/(2\alpha) + \sum_{j=1}^{3} r_{1j}^{P_2\,'}/(3\alpha)$;
      $\varpi_{P_2^2}^{T_1} \leftarrow 1/3 - 1/(2\alpha) + \sum_{j=1}^{3} r_{2j}^{P_2\,'}/(3\alpha)$;
      $\varpi_{P_2^3}^{T_1} \leftarrow 1/3 - 1/(2\alpha) + \sum_{j=1}^{3} r_{3j}^{P_2\,'}/(3\alpha)$;
$t \leftarrow 1$;
Repeat $t++$
      $\boldsymbol{\varpi} \leftarrow (\varpi_{P_1^1}^{T_1} \cdot \varpi_{T_1}^{F}, \varpi_{P_1^2}^{T_1} \cdot \varpi_{T_1}^{F}, \varpi_{P_2^1}^{T_2} \cdot \varpi_{T_2}^{F}, \varpi_{P_2^2}^{T_2} \cdot \varpi_{T_2}^{F}, \varpi_{P_2^3}^{T_2} \cdot \varpi_{T_2}^{F})$;
      $b_i^t \leftarrow \vee_{j=1}^{5} (\varpi_j \wedge \beta_{ji}^t)$, for $1 \le i \le 5$;
      $i_{\max} \leftarrow$ find_max_index($\mathbf{B}^t$);
      $\text{CS}^t \leftarrow y_{i_{\max}}$;
until $t > T_{\max}$
```

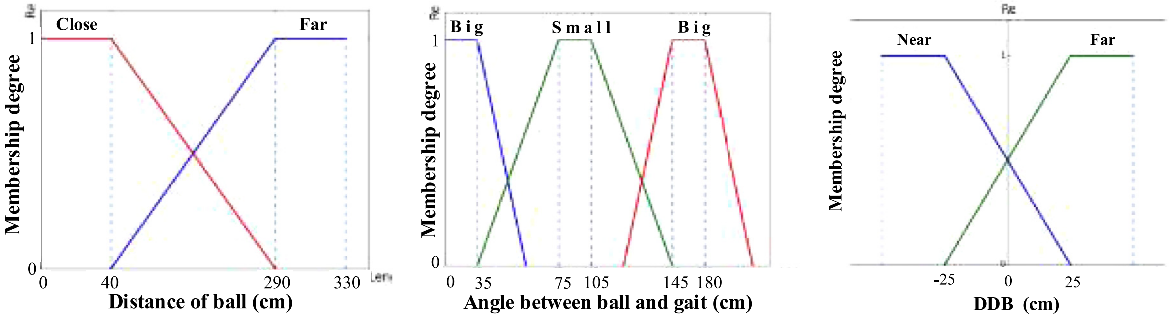


Fig. 3. Membership functions.

4.1. Situation forecast using Bayesian network

4.1.1. Fuzzifier of node information

To construct the Bayesian network, the competition situation (CS), the ball position (BP), and the difference in distance to the ball between us and the opponent (DDB) are used as the parent nodes.

1. CS. The current competition situation can be obtained using FCSEA and it has five fuzzy descriptions: {Superiority, Advantageous, Equilibrium, Disadvantageous, and Weakness}.
2. BP. Ball position is expressed as the threat level for us and it has two fuzzy descriptions: {Advantageous, Disadvantageous}. Judgment for the state of ball position is made according to the distance d from the ball to the opponent's goal and the angle α between the ball and the opponent's goal. Assuming that the current ball's coordinate is (x_b^t, y_b^t) and the goal's coordinate is (x_g, y_g), the current d^t and α^t can be calculated as follows:

$$\begin{cases} d^t = x_b^t - x_g \\ \alpha^t = y_b^t - y_g \end{cases} \tag{13}$$

The distance can be described fuzzily as {Close, Far} according to the membership function shown in Fig. 3(a). The angle can be described fuzzily as {Large, Small} also according to the membership function shown in Fig. 3(b).

3. DDB. The value of current DDB can be calculated as follows:

$$DDB^t = \sum_{i=1}^{n_r} \sqrt{({}^m x_i^t - x_b^t)^2 + ({}^m y_i^t - y_b^t)^2} - \sum_{i=1}^{n_r} \sqrt{({}^o x_i^t - x_b^t)^2 + ({}^o y_i^t - y_b^t)^2} \tag{14}$$

Then DDB^t is fuzzified into two fuzzy descriptions {Near, Far} according to the membership function shown in Fig. 3(c).

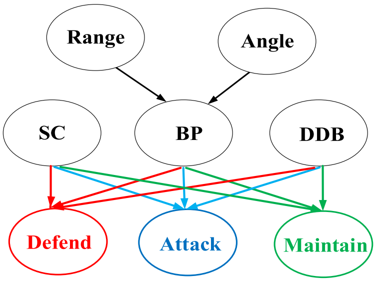


Fig. 4. Diagram of the constructed Bayesian network.

4.1.2. State setting using Bayesian network

According to the actual relationship, a directed acyclic graph, i.e., the Bayesian network is constructed as shown in Fig. 4. The outputs are the independent decision bases expressed as {Attack, Defend, Maintain}. Then, according to the probability of three decision bases, an RL system is executed.

In a Bayesian network [16,13], the child nodes will be affected by their parent nodes. Therefore, according to the rule of probability chain, the Bayesian joint probability for n_v variables is expressed as

$$P(v_1, v_2, ..., v_{n_v}) = \prod_{i=1}^{n_v} P(v_i | g(v_i)) \tag{15}$$

where $g(v_i)$ is the parent node set of v_i.

Therefore, with the current fuzzy description for each root node $\{SC^t, DDB^t, d^t, \alpha^t\}$, the final probabilities are as follows:

$$P(Attack, SC^t, d^t, \alpha^t, DDB^t) = \sum_{i=1}^{2} P(Attack|SC^t, BP_i, DDB^t)P(BP_i|d^t, \alpha^t)P(d^t)P(\alpha^t) \tag{16}$$

$$P(Defend, SC^t, d^t, \alpha^t, DDB^t) = \sum_{i=1}^{2} P(Defend|SC^t, BP_i, DDB^t)P(BP_i|d^t, \alpha^t)P(d^t)P(\alpha^t) \tag{17}$$

$$P(Maintain, SC^t, d^t, \alpha^t, DDB^t) = \sum_{i=1}^{2} P(Maintain|SC^t, BP_i, DDB^t)P(BP_i|d^t, \alpha^t)P(d^t)P(\alpha^t) \tag{18}$$

where BP_i corresponds the two fuzzy descriptions of BP. The probabilities on the right side of (16)-(18) can be obtained using the fuzzy statistical method with data obtained from the $(t-T)^{th}$ time slice to the t^{th} time slice. When $t<T$, data should be collected from the 0^{th} time slice to the t^{th} time slice.

Assuming that the probabilities of three decision bases {Attack, Defend, Maintain} in a period of time, are defined as $\{P_1, P_2, P_3\}$ respectively, each probability is discretized into n_s levels uniformly. For example, for P_1, i.e., the first dimension of state space, the discretization is

$$State_1^t = \begin{cases} 1, 0 \le p_1^t < 1/n_s \\ 2, 1/n_s \le p_1^t < 2/n_s \\ \vdots \\ n_s, (n_s-1)/n_s \le p_1^t < 1 \end{cases} \tag{19}$$

where p_1^t is the current probability calculated using (16), and $State_1^t$ is the current level of p_1^t.

4.2. Adaptive decision-making algorithm using RL (ADMA-RL)

4.2.1. Action set

Since there are usually huge state and action spaces in Markov decision processes, function approximation is a feasible method for dealing with this problem [35]. In addition, an improved dynamic programming method with a goal network is proposed to approximate the partial derivatives of the value function with respect to the system states [38,22]. Nevertheless, dynamic programming requires full information about the model, which is usually not available. Therefore, a data-based optimal control is implemented in [17]. This approach requires only reduced information of measures available at the system outputs. Discretization is also a simple and effective scheme for tackling MDP problems [31]. Therefore, state and action spaces are discretized in the proposed RL method.

According to Section 3.1, there will be n_s^3 states and $\mathbf{s} = (State_1^t, State_2^t, State_3^t)$. For efficient decision-making, the action sets for every state are the same, and every action is a strategy. Assume that there are n_a strategies $\{a_i | a_i = decision\_scheme\_i; i = 1, ..., n_a\}$.

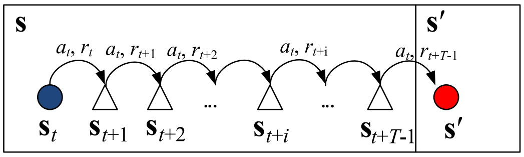


Fig. 5. Semi-Markov diagram.

4.2.2. Reward function

The reward function is categorized into three classes: obtaining a score, losing a score, and normal reward.

1. If our team gets a score, a very good feedback should be given.
2. If our team loses a score, a very bad feedback should be given.
3. The other situations are the normal reward, whose feedback is relative to the current evaluated competition situation by Section 3.

Therefore, the corresponding reward function is

$$r = \begin{cases} 100, & \text{if } \textit{get score} \\ -100, & \text{if } \textit{lose score} \\ 50, & \text{if } CS = Superiority \\ 25, & \text{if } CS = Advantageous \\ 0, & \text{if } CS = Equilibrium \\ -25, & \text{if } CS = Disadvantageous \\ -50, & \text{if } CS = Weakness \end{cases} \tag{20}$$

4.2.3. Action value update

A learning epoch is defined as the time cost of RL for transforming one state into the next state. It may appear that in an epoch, an agent may take the same actions until the current state is transformed into the next state. This is called a semi-Markov phenomenon [21], as shown in Fig. 5. Hence, different from the traditional Q-Value updating strategy, the action taken is

$$Q_{t+1}(\mathbf{s}_t, a_t) = Q_t(\mathbf{s}_t, a_t) + \alpha(r + \gamma \max_{\mathbf{a}_{t+1}} Q_t(\mathbf{s}_{t+1}, a_{t+1})) \tag{21}$$

where s_t represents the state in the t^{th} time slice; α is the learning rate; γ is the discount rate, and s_{t+1} is the next state of s_t after taking a_t.

In the ADMA-RL, the decision-making process is treated as a semi-Markov phenomenon. This is because, in order for an agent to appear in an epoch, the agent may take the same actions until the current state is transformed into the next state. Thus, the Q-Value updating strategy in this paper is different from the traditional Q-Value updating strategy, which is shown in formulas 22, 23 and 24. with reference to [24,18] the convergence of the proposed method can be proved. As for the computational complexity of the proposed method, the size of state space is n_s^3 and the size of action space is n_a. Therefore, the size of policy space is $n_a{}^{n_s^3}$. The training time and computational complexity are satisfied according to the requirements of the experiment.

Assuming that in one epoch, the state s is transformed into s' after λ learning cycle by the same actions shown in Fig. 5, the action value updating strategy in this study is designed as

$$Q_{t+\lambda-1}(\mathbf{s}, a) = r_{t+\lambda-1} + \gamma \max_{\mathbf{a}'} Q(\mathbf{s}', a') \tag{22}$$

$$Q_{t+i}(\mathbf{s}, a) = r_{t+i} + \gamma Q_{t+i+1}(\mathbf{s}, a), 0 \le i \le \lambda - 2 \tag{23}$$

$$Q_{t+\lambda}(\mathbf{s}, a) = Q_t(\mathbf{s}, a) + \alpha(\sum_{k=0}^{\lambda-1} Q_{t+k}/\lambda - Q_t(\mathbf{s}, a)) \tag{24}$$

Then, the learning system can be executed and the corresponding ADMA-RL is shown in Algorithm 2.

5. Experiments and analysis

In this section, the forecast accuracies of decision bases are calculated using the proposed fuzzy Bayesian network (FBN) and the fuzzy neural network (FNN) [14]. The performances are compared with those of the Bayesian SOM neural network (B-SOM) [27] to demonstrate that FBN has better ability in situation forecasting. To demonstrate the practicality and efficiency of RL, the proposed AMDA-RL is tested in the robot soccer platform. Furthermore, to demonstrate the efficiency the following comparisons have been made: 1) the outcomes of the proposed decision-making method and those obtained

Algorithm 2 ADMA-RL.

```
Definition
 $p_i^t$: = probability of $i^{th}$ decision bases in a period of time
 $\mathbf{s}_t$: = current state calculated with $p_i^t$
 $a_t$: = current selected decision scheme with maximum Q-value
 $T_{\max}$ : = maximum competition time
 discretizing_state_space(): = state value calculated with $p_i^t$
 max(): = finding the action with maximum Q-value
 $t \leftarrow 1$;
Repeat $t++$
    $SC^t \leftarrow$ obtained by FCSEA;
    $d^t \leftarrow$ calculated using (12) and blurring;
    $\alpha^t \leftarrow$ calculated using (12) and blurring;
    $DDB^t \leftarrow$ calculated using (13) and blurring;
    $p_1^t \leftarrow \sum_{i=1}^{2} P(Attack|SC^t, BP_i, DDB^t)P(BP_i|d^t, \alpha^t)P(d^t)P(\alpha^t)$;
    $p_2^t \leftarrow \sum_{i=1}^{2} P(Defend|SC^t, BP_i, DDB^t)P(BP_i|d^t, \alpha^t)P(d^t)P(\alpha^t)$;
    $p_3^t \leftarrow \sum_{i=1}^{2} P(Maintain|SC^t, BP_i, DDB^t)P(BP_i|d^t, \alpha^t)P(d^t)P(\alpha^t)$;
    $\mathbf{s}_t \leftarrow$ discretizing_state_space($p_1^t$,$p_2^t$,$p_3^t$);
    $r \leftarrow$ reward_function();
    $a_t \leftarrow$ max(Q(s,a));
    updating Q-value by Section 4.2.3;
until $t > T_{\max}$
```

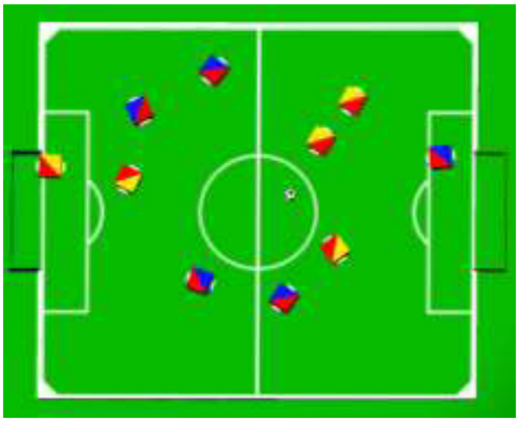


Fig. 6. Simulation platform.

using fuzzy Bayesian reinforcement learning (DM-FBRL), and 2) the decision-making method using fuzzy neural network (DM-FNN), and that using the Bayesian SOM neural network (DM-BSOM).

The experiments are implemented on a robot soccer simulation platform in Webots and there are 10 two-wheeled robots with five robots on our team and five robots on the opponent's team, as shown in Fig. 6. The size of this platform is 330 cm * 180 cm, and the time of each competition is $T_{\max} = 400s$. A supervisor serves as the referee, and it counts the goals and displays the current score and the remaining time. When a team kicks the ball into the goal, the team will get one point. It is especially important that the simulated objects have the same dynamic and physical characteristics as the real objects, such as the robots, the ball, and the ground. Through the simulation platform real-time information, such as the positions and the orientations of all the robots or ball can be obtained, which is convenient for the subsequent analysis. During the test, t is proceeded with three decision-making models: TA (more aggressive attack), TD (stronger defensive ability), TS (the standard model which balances attack and defense).

5.1. Comparison of situation forecast methods

Before the comparison, specific values of each parameter in our system are given as: $\varepsilon_\varphi = \varepsilon_c = \varepsilon_u = 1$, $T = 20s$, $r_c = 10cm$, $n_r = 5$, and $\kappa = (n-1)/2$. The importance matrixes, $R^T = [r_{ij}^T]_{2\times2}$, $R^{P_1} = [r_{ij}^{P_1}]_{2\times2}$, and $R^{P_2} = [r_{ij}^{P_2}]_{3\times3}$, can be obtained through a number of competitions, and the results are shown in Table 1 (Table 1.1., Table 1.2., Table 1.3.).

According to Table 1 and (7)-(9), the final weights for evaluation factors $\{\eta^t, S^t, \varphi^t, c^t, u^t\}$ can be calculated as follows:

$$\varpi = (\varpi_T^{P_1} \cdot \varpi_{P_1}^1, \varpi_T^{P_1} \cdot \varpi_P^2, \varpi_T^{P_2} \cdot \varpi_{P_2}^1, \varpi_T^{P_2} \cdot \varpi_{P_2}^2, \varpi_T^{P_2} \cdot \varpi_{P_2}^3) = (0.27, 0.33, 0.1133, 0.1533, 0.1334) \tag{25}$$

Table 1.1
Precedence relations of *F*-*T*.

	T_1	T_2
T_1	0.50	0.70
T_2	0.30	0.50

Table 1.2
Precedence relations of T_1-P_1.

	${P_1}^1$	${P_1}^2$
${P_1}^1$	0.50	0.40
${P_1}^2$	0.60	0.50

Table 1.3
Precedence relations of T_2-P_2.

	${P_2}^1$	${P_2}^2$	${P_2}^3$
${P_2}^1$	0.50	0.30	0.40
${P_2}^2$	0.70	0.50	0.60
${P_2}^3$	0.60	0.40	0.50

Table 2
Fuzzy membership for evaluation factors.

	y_1	y_2	y_3	y_4	y_5
$\eta^t \in [0, 60)$	0.00	0.00	0.10	0.80	1.00
$\eta^t \in [60, 180)$	0.00	0.10	0.40	0.80	0.40
$\eta^t \in [180, 300)$	0.10	0.20	0.60	0.60	0.20
$\eta^t \in [300, 480)$	0.60	0.80	0.50	0.10	0.00
$\eta^t \in [480, 600)$	1.00	0.80	0.10	0.00	0.00
$S^t \in [-\infty, -6)$	0.00	0.00	0.00	0.80	1.00
$S^t \in [-5, \ -2]$	0.00	0.00	0.10	0.80	0.40
$S^t \in [-1, 1]$	0.10	0.20	0.60	0.20	0.10
$S^t \in [2, 5]$	0.60	0.80	0.40	0.10	0.00
$S^t \in [6, +\infty)$	1.00	0.80	0.00	0.00	0.00
$\varphi^t \in [0, 0.4)$	0.20	0.20	0.40	0.80	0.40
$\varphi^t \in [0.4, 0.6)$	0.40	0.80	0.80	0.20	0.20
$\varphi^t \in [0.6, 1)$	0.60	0.80	0.60	0.40	0.10
$c^t \in [0, 0.4)$	0.20	0.20	0.40	0.80	0.40
$c^t \in [0.4, 0.6)$	0.40	0.80	0.80	0.20	0.20
$c^t \in [0.6, 1)$	0.60	0.80	0.60	0.40	0.10
$u^t \in [0, 0.4)$	0.10	0.20	0.40	0.80	0.60
$u^t \in [0.4, 0.6)$	0.40	0.60	0.80	0.50	0.40
$u^t \in [0.6, 1)$	0.60	0.80	0.50	0.40	0.30

Through a large number of competitions and the fuzzy statistical method, the fuzzy membership table for evaluation factors can be obtained, as shown in Table 2.

To compare the situation forecast, the three methods of, FBN, FNN, and B-SOM, are tested using the supervised data. In particular, for FBN, when the data of one evaluation factor are selected, the corresponding fuzzy membership vector for the situations **Y** can be obtained by checking Table 2. For example, if c^t is 0.5, the evaluation vector {0.40, 0.80, 0.80, 0.20, 0.20} of c^t can be obtained by checking Table 2. The evaluation vector of other factors can also be obtained in a similar way. These vectors constitute fuzzy relation matrix β^t. Then, with the obtained real-time fuzzy relation matrix β^t, the final comprehensive competition situation can be calculated using (12). Outputs of the Bayesian network are the probabilities of decision bases, Attack, Defend, and Maintain; hence, the predicted decision base is the description with maximum probability. The accuracies, i.e. the ratios of a correct forecast, of the three methods, are shown in Fig. 7. In Fig. 7, the x axis denotes number of iterations and the y axis denotes accuracy rate.

In FNN and FBN, knowledge representation is expressed in fuzzy logic, which is more similar to human thinking. Its knowledge representation ability is better than the Bayesian method, while the precisions of FNN and FBN are good with less iterations. FNN and FBN rely too much on expert experience. With more iterations, fuzzy logic accuracy is convergent

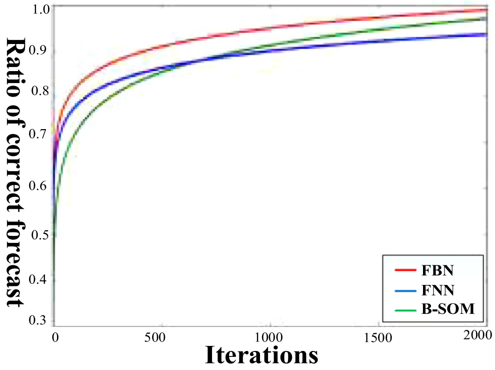
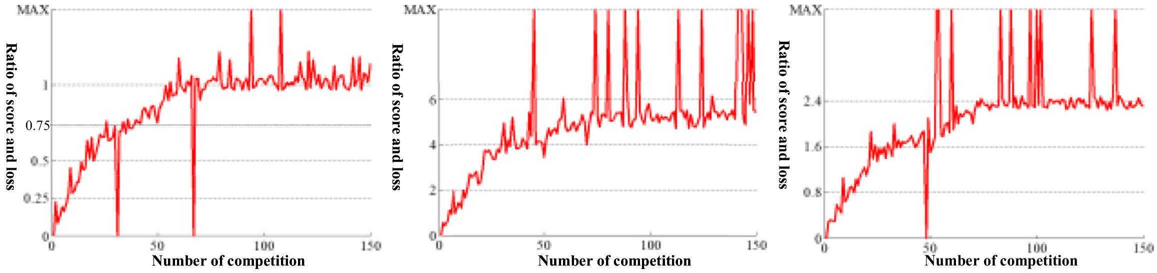


Fig. 7. Comparison of results obtained by the three situation forecast methods.



(a) Ratio change of DM-FBRL and TA

(b) Ratio change of DM-FBRL and TD

(c) Ratio change of DM-FBRL and TS

Fig. 8. Ratios over competitions.

Table 3
Intrinsic parameters of RL.

Identified Parameter	VALUE
n_s	3
n_a	5
γ	0.85
α	0.8

but lower than that of Bayesian inference. The proposed method combines the superior aspects of fuzzy logic and the Bayesian method, and it, therefore, obtains smaller mean square error (MSE) from the beginning to the end.

5.2. Test results of ADMA-RL

To demonstrate the efficiency of the proposed ADMA-RL, TA, TD, and TS are used as the opponents for learning. Each competition lasts 400 s, and the ratios of scores and losses over the competitions are recorded as shown in Fig. 8. The ratio is calculated as score/loss. When loss is 0, “MAX” denotes the radio in Fig. 8. The intrinsic parameters of RL are shown in Table 3. In Fig. 8, the x axis denotes number of competitions and y axis denotes ratio of score and loss.

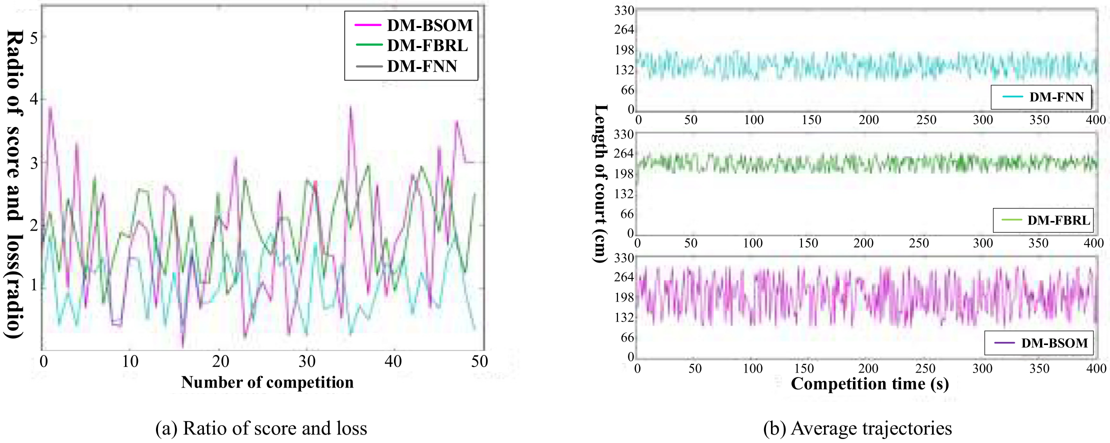
As seen in Fig. 8, when ADMA-RL competes with TA, the ratio rises gradually and stabilizes after 65 competitions. When ADMA-RL competes with TD, the ratio rises gradually and stabilizes after 60 competitions. When ADMA-RL competes with TS, the ratio rises gradually and stabilizes after 75 competitions. Thus, it can be concluded that along with continuous learning, ADMA-RL can become increasingly intelligent and scores more in one competition, demonstrating its efficiency.

5.3. Comparison of dynamic decision-making methods

The performance of DM-FBRL, DM-FNN, and DM-BSOM are evaluated through 50 robot soccer games and each competition lasts 400 s with the TS model. The results shown in Table 4 and Fig. 9, are the score ratio of each match and the effect of strategy with that model, respectively. Fig. 9(a) shows Ratio of score and loss. In Fig. 9(a), the x axis denotes number of competitions and the y axis denotes the Ratio of score and loss. Fig. 9(b) shows the average trajectories of a ball when it scores with the min-max coordinate filtered out. In Fig. 9(b), the x axis denotes time and the y axis denotes the ball’s abscissas position. In robot soccer the ball position shows the team’s competitiveness and defense ability.

Table 4
Confrontation result.

Confrontation factories	DM-FNN	DM-BSOM	DM-FBRL
Total ball-controlling rate	43.3%	58.7%	66.2%
Threat degree to gate (goal times/shooting times)	22/871	71/1489	44/1102
Total number of losing ball	17	52	29



(a) Ratio of score and loss

(b) Average trajectories

Fig. 9. Score ratio and average ball trajectories.

As seen in Table 4 and Fig. 9, the decision-making results of DM-FBRL, DM-FNN, and DM-BSOM are better than those of the TS model. As seen in Fig. 9(a), the ratio of score and loss for DM-BSOM fluctuates greatly, ranging from 0 to 4. The ratios of score and loss for the remaining methods have low fluctuations. The ratio of the score and loss of DM-FBRL fluctuates around 2 while that of DM-FNN fluctuates around 1. As seen in Fig. 9(b), the average trajectories of DM-BSOM have a large range, approximately from 100 to 300. The average trajectories of DM-FBRL changes from 200 to 264 and that of DM-FNN varies from100 to 200, indicating that the DM-FNN model is over-reliant on expert experience and thus cannot accurately predict situations. DM-FNN is always in a defensive state with only limited time to control the ball or choose an effective game strategy during a confrontation. That leads to a lower score ratio and less aggressive behavior towards opponents. In contrast, DM-BSOM and DM-FBRL can develop a basic confrontation strategy according to the situations during a game and can control the ball for longer periods of time. DM-BSOM is more aggressive and it can get more scores. However, DM-BSOM frequently changes to different attack and defense strategies, which may lead to poorer stability and losing more games than DM-FBRL. Compared with the other two methods, DM-FBRL has the advantages of more appropriate situation evaluation and more stable strategy selection. Therefore, the DM-FBRL model can predict the game situations more effectively and form strategies more precisely.

In addition, the accuracy of situation evaluation is validated with DM-FBRL and DM-BSOM used in the robot soccer games with TA and TD methods. They are in a disadvantaged state when competing with TA, but they have overall advantages when competing with TD. Fig. 10 shows the ball trajectories in the confrontation. In Fig. 10, the x axis denotes competition time and the y axis denotes the ball's abscissas position.

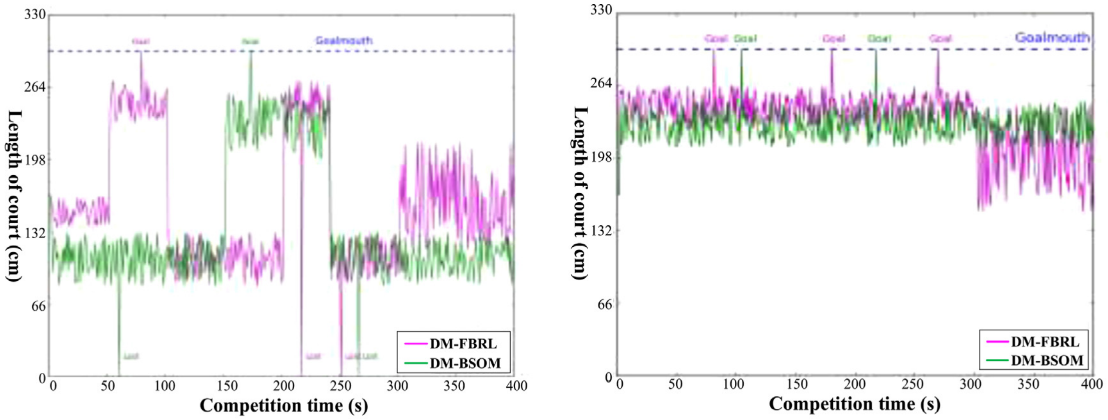
When competing with TA, DM-BSOM scores in 173 s and loses in {62 s, 274 s}. DM-FBRL scores in 92 s, loses in {127 s, 271 s}, as shown in Fig. 10(a). However, when the competition is almost over (match time is 300 s), the TA model is leading by 2:1. DM-BSOM still selects a defensive strategy, whereas the DM-FBRL selects a more aggressive strategy to pursue the score, even though it may not be successful; meaning that DM-FBRL can represent better expert knowledge.

In Fig. 10(b) DM-BSOM and DM-FBRL have absolute advantages in their match. After 300 s of the game, DM-FBRL prefers to choose a conservative formation to control the pace. That can avoid accidentally losing a score, and this is essential to winning a game in robot soccer. On the contrary, an overly aggressive strategy, such as that adopted by DM-BSOM, may lose even in a very advantageous position because of its poor stability. Thus, DM-FBRL is a better choice in robot soccer.

In comparison, the DM-FBRL model combines the knowledge representation of fuzzy logic with the inferential capability of Bayesian logic. After evaluating a situation, the model can accurately infer the robot soccer situation and choose the right strategy formation. This procedure is more similar to human thinking.

6. Conclusions

In view of the problems of complex models and uncertain knowledge representation in dynamic decision-making methods, this study proposes an adaptive decision-making method that involves fuzzy Bayesian reinforcement learning. Sequen-



(a) Confrontation with TA

(b) Confrontation with TD

Fig. 10. Ball trajectories.

tially, a fuzzy situation evaluation proceeds first to access the situation. Then, a hierarchy with real-time and sequential situations is combined with the FCEM. Since uncertain knowledge reduces the inference ability of the fuzzy evaluation method, a Bayesian network is designed. Both inputs and outputs of the network are the comprehensively evaluated competition situations and the bases for decision. According to the probabilities of decision bases, the RL method is utilized to learn an appropriate policy. Experimental results demonstrate that the proposed method can provide good knowledge representation and efficient strategy selection. The contributions of the proposed method are as follows.1. Efficient situation evaluation; the fuzzy situation evaluation method combines expert experience with the actual game. 2. Adaptive policy learning; the RL approach can allow learning to be more effective because the agent can learn an effective strategy with RL in a different environment. Through the RL process, the appropriate strategy can be mapped to the state. 3. Flexibility in state aggregation; the fuzzy method is applied to state classification. If there are too many or too few states, the membership functions or the discretization factor n_s can be adjusted to change the numbers of states as well as the classification of states.

Acknowledgement

The research work was supported by National research and development plan of China (No. 2017YFB1001900), National Natural Science Foundation of Shaanxi (No. 2015JM6308), the fund of national ministries (No. 2016ZC53022), and the Fundamental Research Funds for the Central Universities(No. Z2017231).

References

[1] X.P. Burgos-Artizzu, A. Ribeiro, A. Tellaeche, et al., Improving weed pressure assessment using digital images from an experience-based reasoning approach, Comput. Electron. Agric. 65 (2009) 176–185.
[2] A.G.S. Conceicao, C.E.T. Dórea, L. Martinez, et al., Design and implementation of model-predictive control with friction compensation on an omnidirectional mobile robot, IEEE/ASME Trans. Mechatron. 19 (2014) 467–476.
[3] B. Chen, A. Zhang, L. Cao, Autonomous intelligent decision-making system based on Bayesian SOM neural network for robot soccer, Neurocomputing 128 (2014) 447–458.
[4] F. Castaldo, F.A.N. Palmieri, C.S. Regazzoni, Bayesian analysis of behaviors and interactions for situation awareness in transportation systems, IEEE Trans. Intell. Transp. Syst. 17 (2016) 313–322.
[5] M. Elkano, M. Galar, J. Sanz, et al., Fuzzy rule-based classification systems for multi-class problems using binary decomposition strategies: on the influence of n-dimensional overlap functions in the fuzzy reasoning method, Inf. Sci. 332 (2016) 94–114.
[6] İ. Ertuğrul, N. Karakaşoğlu, Performance evaluation of Turkish cement firms with fuzzy analytic hierarchy process and TOPSIS methods, Expert Syst. Appl. 36 (2009) 702–715.
[7] Y. Hu, Y. Gao, B. An, Accelerating multiagent reinforcement learning by equilibrium transfer, IEEE Trans. Cybern. 45 (2015) 1289–1302.
[8] K.S. Hwang, S.W. Tan, C.C. Chen, Cooperative strategy based on adaptive Q-learning for robot soccer systems, Fuzzy Syst. IEEE Trans. 12 (2004) 569–576.
[9] A. Joseph, N.E. Fenton, M. Neil, Predicting football results using Bayesian nets and other machine learning techniques, Knowl.-Based Syst. 19 (2006) 544–553.
[10] E.R. Jalao, T. Wu, D. Shunk, A stochastic AHP decision making methodology for imprecise preferences, Inf. Sci. 270 (2014) 192–203.
[11] C.J. Kim, D. Chwa, Obstacle avoidance method for wheeled mobile robots using interval type-2 fuzzy neural network, IEEE Trans. Fuzzy Syst. 23 (2015) 677–687.
[12] R. Kumar, B. Singh, D.T. Shahani, et al., Recognition of power-quality disturbances using S-transform-based ANN classifier and rule-based decision tree, IEEE Trans. Ind. Appl. 51 (2015) 1249–1258.
[13] B.U. Kim, D. Goodman, M. Li, et al., Improved reliability-based decision support methodology applicable in system-level failure diagnosis and prognosis, IEEE Trans. Aerosp. Electron. Syst. 50 (2014) 2630–2641.
[14] S.H. Kasaei, S.M. Kasaei, S.A. Kasaei, et al., Dynamic role engine and formation control for cooperating agents with robust decision-making algorithm, Ind. Rob.: Int. J. 38 (2011) 153–162.
[15] J. Lu, V. Behbood, P. Hao, et al., Transfer learning using computational intelligence: a survey, Knowl.-Based Syst. 80 (2015) 14–23.
[16] P. Larrañaga, H. Karshenas, C. Bielza, et al., A review on evolutionary algorithms in Bayesian network learning and inference tasks, Inf. Sci. 233 (2013) 109–125.

[17] F.L. Lewis, K.G. Vamvoudakis, Reinforcement learning for partially observable dynamic processes: adaptive dynamic programming using measured output data, IEEE Trans. Syst. Man Cybern. B Cybern. Publ. IEEE Syst. Man Cybern. Soc. 41 (2011) 14–25.
[18] N. Limnios, G. Oprisan, Semi-Markov Processes and Reliability, Springer Science & Business Media, 2012.
[19] A.T. Misirli, A.B. Bener, Bayesian networks for evidence-based decision-making in software engineering, IEEE Trans. Softw. Eng. 40 (2014) 533–554.
[20] S. Mohagheghi, Integrity assessment scheme for situational awareness in utility automation systems, IEEE Trans. Smart Grid 5 (2014) 592–601.
[21] N. Marchenko, C. Bettstetter, Cooperative ARQ with relay selection: an analytical framework using semi-Markov processes, IEEE Trans. Veh. Technol. 63 (2014) 178–190.
[22] Z. Ni, H. He, D. Zhao, et al., GrDHP: a general utility function representation for dual heuristic dynamic programming, IEEE Trans. Neural Netw. Learn. Syst. 26 (2015) 614–627.
[23] D. Ozgen, B. Gulsun, Combining possibilistic linear programming and fuzzy AHP for solving the multi-objective capacitated multi-facility location problem, Inf. Sci. 268 (2014) 185–201.
[24] R.E. Parr, Hierarchical Control and Learning For Markov decision Processes, University of California at Berkeley, 1998.
[25] R.M. RodríGuez, L. MartıNez, F. Herrera, A group decision making model dealing with comparative linguistic expressions based on hesitant fuzzy linguistic term sets, Inf. Sci. 241 (2013) 28–42.
[26] M. Riedmiller, T. Gabel, R. Hafner, et al., Reinforcement learning for robot soccer, Auton. Rob. 27 (2009) 55–73.
[27] S. Ribaric, T. Hrkac, A model of fuzzy spatio-temporal knowledge representation and reasoning based on high-level Petri nets, Inf. Syst. 37 (2012) 238–256.
[28] H. Shi, L. Xu, L. Zhang, et al., Research on self-adaptive decision-making mechanism for competition strategies in robot soccer, Front. Comput. Sci. 9 (2015) 485–494.
[29] H. Shi, Z. Yu, Y. Xu, et al., The study of situation evaluation in SimuroSot decision support systems, Kybernetes 41 (2012) 1226–1234.
[30] H. Shi, L. Zhang, W. Pan, et al., Robot soccer confrontation decision-making technology based on MOGM: multi-objective game model, J. Intell. Fuzzy Syst. 28 (2015) 713–724.
[31] H. Shi, X. Li, K.S. Hwang, et al., Decoupled Visual Servoing with Fuzzy Q-Learning, IEEE Trans. Ind. Inf. 14 (2018) 241–252.
[32] W. Tao, G. Zhang, Trusted interaction approach for dynamic service selection using multi-criteria decision making technique, Knowl.-Based Syst. 32 (2012) 116–122.
[33] J. Wang, J.Q. Wang, H.Y. Zhang, et al., Multi-criteria decision-making based on hesitant fuzzy linguistic term sets: an outranking approach, Knowl.-Based Syst. 86 (2015) 224–236.
[34] X. Wei, X. Luo, Q. Li, et al., Online comment-based hotel quality automatic assessment using improved fuzzy comprehensive evaluation and fuzzy cognitive map, IEEE Trans. Fuzzy Syst. 23 (2015) 72–84.
[35] F.Y. Wang, H. Zhang, D. Liu, Adaptive dynamic programming: an introduction, IEEE Comput. Intell. Mag. 4 (2009) 39–47.
[36] C.C. Wong, C.T. Cheng, H.M. Chan, Design and implementation of an autonomous robot soccer system, Int. J. Adv. Rob. Syst. 10 (2013) 1–13.
[37] Z. Xu, K. Gao, T.M. Khoshgoftaar, et al., System regression test planning with a fuzzy expert system, Inf. Sci. 259 (2014) 532–543.
[38] X. Xu, Z. Huang, L. Zuo, Reinforcement learning algorithms with function approximation: recent advances and applications, Inf. Sci. 261 (2014) 1–31.
[39] Z. Zhang, Hesitant fuzzy power aggregation operators and their application to multiple attribute group decision making, Inf. Sci. 234 (2013) 150–181.
[40] M. Zolfpour-Arokhlo, A. Selamat, S.Z.M. Hashim, et al., Modeling of route planning system based on Q value-based dynamic programming with multi--agent reinforcement learning algorithms, Eng. Appl. Artif. Intell. 29 (2014) 163–177.