Applied Soft Computing Journal 102 (2021) 107109

Contents lists available at ScienceDirect

Applied Soft Computing Journal

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

Data-driven preference learning in multiple criteria decision making in the evidential reasoning context

Chao Fu [a,b,c], Min Xue [a,b,c,*], Weiyong Liu [d], Dongling Xu [e], Jianbo Yang [e]

[a] School of Management, Hefei University of Technology, Hefei, Box 270, Hefei 230009, Anhui, PR China
[b] Key Laboratory of Process Optimization and Intelligent Decision-making, Ministry of Education, Hefei 230009, Anhui, PR China
[c] Ministry of Education Engineering Research Center for Intelligent Decision-Making & Information System Technologies, Hefei 230009, Anhui, PR China
[d] Department of Ultrasound, The First Affiliated Hospital of USTC, Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei 230036, Anhui, PR China
[e] Manchester Business School, The University of Manchester, Manchester M15 6PB, UK

ARTICLE INFO

Article history:
Received 18 August 2020
Received in revised form 12 December 2020
Accepted 12 January 2021
Available online 19 January 2021

Keywords:
Data-driven preference learning
Multiple criteria decision making
Evidential reasoning approach
Evidential reasoning rule
Diagnosis of thyroid nodules

ABSTRACT

In some situations, such as the diagnosis of thyroid nodules, a decision maker considers observations on multiple criteria to provide the overall assessments and advice on what will be done in the next step. To guarantee the quality of the assessments and advice and their consistency with observations, this paper proposes a method of learning the preferences of the decision maker from the observations on multiple criteria and the overall assessments provided. The constraints on preferences are learned first to avoid extreme and irrational preferences. Within the feasible region formed by the constraints, the preferences are learned. When gold standards, which can be used to judge the correctness of the overall assessments, are available, the issue of how to learn the constraints and the preferences that satisfy the constraints is presented. With and without the consideration of gold standards, the way in which solutions can be generated using the learned preferences is introduced. To demonstrate the process of preference learning based on observations and overall assessments, a case study is conducted using the examination reports generated by three radiologists from 2013 to 2017 in a hospital located in Hefei, Anhui, China.

1. Introduction

In daily life, people often face the problems of selecting or ranking objects (or alternatives) from multiple perspectives (or criteria), such as ranking MBA programs, schools, or universities [1], selecting interesting websites [2], selecting a summer holiday destination [3], or choosing a film to watch [3]. This process is referred to as multiple criteria decision making (MCDM) [4–7].

With a view to facilitating the analysis of real-world MCDM problems, different types of traditional methods have been proposed to show great performance when the assessments on each criterion and the relevant decision parameters, such as criterion weights, can be provided by a decision maker [8–12]. In a new era of Internet and big data, however, large amounts of data lead to opportunities and challenges that coexist for MCDM. On the one hand, the available data allow a decision maker to make assessments more objectively to avoid possible subjective bias. On the other hand, the algorithms or tools to estimate the decision maker's preferences from the data are required so that satisfactory and rational solutions to the problem can be obtained. Under these conditions, preference learning has become an important issue, which is the scientific problem of this paper. Many attempts related to the problem have been made in existing studies.

In machine learning, preference learning is the induction of preference models from observed data that characterize decision makers' preferences and then the use of the models to predict the decision makers' preferences in new problems [1,13,14]. Two modes of preference structures are usually learned or estimated from the available data, which include value (or utility) functions and binary preference relations [13,14]. The predicted decision makers' preferences are anticipated to be as close to the real ones as possible. For this purpose, the predicted preferences dynamically change with an increase in the available data on the decision makers' choices or evaluations. In recommender systems, preference learning is the creation, storage, and dynamic updating of decision makers' preference profiles from their implicit or explicit feedback as well as the prediction of the decision makers' future

* Corresponding author at: School of Management, Hefei University of Technology, Hefei, Box 270, Hefei 230009, Anhui, PR China.
E-mail address: xuehfut@163.com (M. Xue).

https://doi.org/10.1016/j.asoc.2021.107109
1568-4946/

selections [3,15–18]. After the decision makers' selections are made, the over-ranked items and the selected items are used to help adapt the decision makers' preference profiles to improve the prediction accuracy in future recommendations [3]. Then, the selected items and the updated preference profiles of the decision makers are used to update the criterion weights for the same purpose [17].

In MCDM, the preferences of a decision maker may be elicited from his or her past decisions in similar situations [1]. Given the pairwise preferences between the pairs of reference alternatives and the marginal evaluations of the alternatives on each criterion, criterion weights are learned by means of maximum likelihood estimation [19]. Different from machine learning and recommender systems, the estimation of marginal preference structures on each criterion is not easy, especially in uncertain contexts. Even in very similar decision situations, a decision maker may provide different marginal evaluations on some or all criteria due to controllable or uncontrollable factors, such as the change in the expertise and experience of the decision maker, the organization environment, and the external environment. In addition, in some decision contexts, there is a one-to-one correspondence between the observation and the assessment of an item. For example, when a radiologist diagnoses whether a thyroid nodule is cancerous from the perspectives of margin, contour, echogenicity, calcification, and vascularity, the observation on each criterion can be directly transformed into the assessment of the nodule on the criterion by using the common expertise and experience of the ultrasonic department in which the radiologist works. In this situation, the preferences of a decision maker are reflected by criterion weights. In consideration of these facts, the preferences of a decision maker in MCDM are characterized by criterion weights in this study.

The purpose of this study is to learn the preferences of a decision maker in MCDM and then use the learned preferences to generate solutions to new MCDM problems when historical decision data are available. For the preferences characterized by criterion weights, historical data are composed of historical assessments on criteria and historical overall assessments. Note that MCDM in this study is not only to select the best option [20–23], but also to help decide what will be done in the next step.

For the above purpose, this paper proposes a method of learning preferences in MCDM in the context of the evidential reasoning (ER) approach [24–26]. On the condition that the assessments on criteria and the overall assessments provided are available, the interval of average similarity between the two types of assessments are determined and used to help specify the constraints on criterion weights. This is done to avoid extreme weights that are inconsistent with the historical assessments on criteria and the historical overall assessments. With the consideration of the constraints, the average difference between the aggregated assessments derived from combining the assessments on criteria through the ER rule [27] and the overall assessments provided is minimized to generate the learned criterion weights. In another situation in which gold standards, which can be used to judge the correctness of the overall assessments, are available, the reliability of a decision maker is measured by the overall assessments and the gold standards. By considering the reliability of the decision maker as his or her learning rate, the constraints on criterion weights and the criterion weights are learned from the assessments on criteria, the overall assessments, and the gold standards. The issue of how to generate solutions with and without the consideration of gold standards is then introduced, in which the effect of the learned criterion weights is examined. To verify the effectiveness of the proposed method, a problem of learning radiologists' preferences from historical examination reports is investigated to help the radiologists provide the consistent diagnoses or improve their diagnostic capability. The historical examination reports and the corresponding pathologic findings as gold standards from 2013 to 2017 associated with three radiologists in the ultrasonic department of a tertiary hospital located in Hefei, Anhui, China are collected. They are used to demonstrate how to learn the preferences of the three radiologists in diagnosing thyroid nodules.

The main contributions of this paper include the following: (1) a method of learning a decision maker's preferences from historical data in MCDM is proposed in the ER context; (2) preference learning based on historical data in MCDM is conducted with and without gold standards; (3) decisions through the learned preferences are made with and without the consideration of gold standards; and (4) the proposed method is used to help radiologists diagnose thyroid nodules, in which the contributions of the proposed method to the diagnosis of nodules are examined and the separation-integration of historical data is discussed by simulation.

The rest of this paper is organized as follows. Section 2 recalls the ER approach. Section 3 presents preference learning in MCDM when the assessments on criteria and the overall assessments are available. Section 4 presents how to make decisions through the learned preferences. A case study on the diagnosis of thyroid nodules is conducted to demonstrate the preference learning and the decision making from the learned preferences in Section 5. Finally, the paper's conclusions are presented in Section 6.

2. Preliminaries

As the method of learning preferences is developed in the context of the ER approach [24–26], the approach is simply recalled in the following.

Through a set of grades $\Omega = \{H_1, H_2, \ldots, H_N\}$ that is increasingly ordered from worst to best, alternative a_l ($l = 1, \ldots, M$) is evaluated on criterion e_i ($i = 1, \ldots, L$) in the ER approach. The utilities of grades $u(H_n)$ ($n = 1, \ldots, N$), which satisfy $0 = u(H_1) < u(H_2) < \ldots < u(H_N) = 1$ are used to reflect the difference among grades. The evaluation is profiled by a belief distribution $B(e_i(a_l))$ = $\{(H_n, \beta_{n,i}(a_l)), n = 1, \ldots, N; (\Omega, \beta_{\Omega,i}(a_l))\}$, where $\beta_{n,i}(a_l)$ with $\beta_{n,i}(a_l) \geq 0$ and $\sum_{n=1}^{N} \beta_{n,i}(a_l) \leq 1$ denotes the belief degree assigned to grade H_n, and $\beta_{\Omega,i}(a_l) = 1 - \sum_{n=1}^{N} \beta_{n,i}(a_l)$ represents the degree of global ignorance [28]. If $\beta_{\Omega,i}(a_l) = 0$, the assessment is complete; otherwise, it is incomplete. The belief distribution of each alternative on each criterion forms a belief decision matrix $S_{L\times M}$.

Suppose that criterion weights are represented by $w = (w_1, w_2, \ldots, w_L)$ such that $0 \leq w_i \leq 1$ and $\sum_{i=1}^{L} w_i = 1$. Through w and the ER rule [27], the assessments $B(e_i(a_l))$ ($i = 1, \ldots, L$, $l = 1, \ldots, M$) are combined to generate the overall assessment $B(a_l) = \{(H_n, \beta_n(a_l)), n = 1, \ldots, N; (\Omega, \beta_\Omega(a_l))\}$ ($l = 1, \ldots, M$). Similar to $\beta_{\Omega,i}(a_l)$, $\beta_\Omega(a_l)$ represents the degree of the aggregated global ignorance. From the overall assessment $B(a_l)$, the minimum and maximum expected utilities of the alternative a_l are calculated as $u^-(a_l)=\sum_{n=2}^{N} \beta_n(a_l)u(H_n)+(\beta_1(a_l)+\beta_\Omega(a_l))u(H_1)$ and $u^+(a_l)=\sum_{n=1}^{N-1} \beta_n(a_l)u(H_n)+(\beta_N(a_l)+\beta_\Omega(a_l))u(H_N)$.

The ER rule [27] is used to implement the combination of assessments on criteria, which is simply presented as follows.

Definition 1 (*[27]*). Given the assessments $B(e_i(a_l))$ ($i = 1, \ldots, L$) and their weights w_i, the combination result of the first i assessments is defined as

$$\{(H_n, \beta_{n,b(i)}(a_l)), n = 1, \ldots, N; (\Omega, \beta_{\Omega,b(i)}(a_l))\}, \tag{1}$$

where, it is satisfied that $0 \leq \beta_{n,b(i)}(a_l)$, $\beta_{\Omega,b(i)}(a_l) \leq 1$, and the detailed explanation of Definition 1 is shown in Appendix A of the supplementary material.

3. Preference learning in MCDM

In this section, how to learn the preferences of a decision maker in MCDM as to criterion weights from historical data is described.

3.1. Preference learning without gold standards

In traditional MCDM with belief distributions, the assessments of alternatives on each criterion are combined by using the ER rule and criterion weights to generate the overall assessments of alternatives. The resulting overall assessments are then used to compare alternatives and generate a solution. In practice, this is not always the case [20–23]. In addition to comparing alternatives, the purpose of MCDM may also be to make judgments on alternatives and provide service to the next decision. For example, by observing the features (as well as the criteria) of an inpatient's nodule and providing diagnostic opinions from the observations, a radiologist is required to provide the overall opinion about the degree to which the nodule is a malignant lesion. A clinician (or surgeon) decides how to handle the inpatient with the consideration of the overall opinion of the radiologist. The two purposes of MCDM are addressed in this paper.

Facing an MCDM problem with belief distributions, a decision maker must provide the assessments of alternatives on each criterion and synthetically consider the assessments to provide the overall assessments of alternatives with the consideration of criterion weights. Unlike traditional MCDM, it is not required that criterion weights are provided by the decision maker. It is not possible to require the decision maker to generate the overall assessment from the assessments on criteria precisely with the aid of criterion weights. However, this does not mean that criterion weights are not important in MCDM. In contrast, criterion weights are important for helping a decision maker to provide the overall assessment from the assessments on criteria. Inspired by the importance of criterion weights, an interesting issue is to determine criterion weights in MCDM on the condition that they are not subjectively specified by a decision maker.

Suppose that a decision maker has handled the same type of MCDM problems in the past, where different alternatives are evaluated by using the same set of grades on a common set of criteria. Under this assumption, a number of historical decision matrices and overall assessments are accumulated. The issue of determining criterion weights is transformed into learning criterion weights from historical decision matrices and overall assessments. The learned weights reflect the preferences of a decision maker because historical decision matrices and overall assessments are provided by the decision maker.

To learn criterion weights from historical decision matrices and overall assessments, constraints on weights are required first to avoid extreme weights. By using extreme weights to generate overall assessments, there are certainly some criteria on which the assessments contribute little or nothing to the overall assessments. This may not be what is anticipated by a decision maker. To obtain criterion weights that are consistent with what is anticipated by a decision maker, constraints on weights are also learned from historical decision matrices and overall assessments. In theory, if the weight on a criterion is large, the contribution of the assessment on the criterion to the overall assessment is large. Without knowing the decision maker's subjective judgments on criterion weights, a higher similarity between the assessment on a criterion and the overall assessment means a larger contribution of the assessment on the criterion to the overall assessment. By following this idea to learn constraints on criterion weights, the similarity between the assessment on a criterion and overall assessment must be measured.

Assume that the overall assessment of alternative a_l provided by a decision maker is represented by $\widetilde{B}(a_l) = \{(H_n, \widetilde{\beta}_n(a_l)), n = 1, \ldots, N; (\Omega, \widetilde{\beta}_{\Omega}(a_l))\}$. To learn constraints on criterion weights from $B(e_i(a_l))$ and $\widetilde{B}(a_l)$, the similarity between them is measured by using the dissimilarity measure between two belief distributions developed by Fu et al. [29]. For simplicity, the situation of complete $B(e_i(a_l))$ and $\widetilde{B}(a_l)$ is considered first.

Definition 2 (*[29]*)**.** Suppose that the distributed dissimilarity between assessments $B(e_i(a_l))$ and $\widetilde{B}(a_l)$ is defined as

$$GD(e_i(a_l)) = \{(H_n, \overline{\beta}_{n,i}(a_l) = \left|\beta_{n,i}(a_l) - \widetilde{\beta}_n(a_l)\right|), n = 1, \ldots, N\}. \tag{2}$$

Then, a dissimilarity measure between the two assessments is constructed using $GD(e_i(a_l))$ as

$$\widetilde{D}(e_i(a_l)) = \sum_{n=1}^{N-1} \sum_{m=n+1}^{N} \overline{\beta}_{n,i}(a_l) \cdot \overline{\beta}_{m,i}(a_l) \cdot (u(H_m) - u(H_n)). \tag{3}$$

In accordance with Definition 2, the similarity between $B(e_i(a_l))$ and $\widetilde{B}(a_l)$ can be measured by

$$\widetilde{S}(e_i(a_l)) = 1 - \widetilde{D}(e_i(a_l)). \tag{4}$$

As demonstrated in [29], both $\widetilde{D}(e_i(a_l))$ and $\widetilde{S}(e_i(a_l))$ are limited to [0, 1]. Consider the union of historical decision matrices as a historical decision matrix with $\overline{M}$ alternatives. From Eqs. (2)–(4), the average similarity between the assessments of alternatives on criterion e_i and the overall assessments of the alternatives is calculated as

$$\widetilde{S}(e_i) = \frac{\sum_{l=1}^{\overline{M}} \widetilde{S}(e_i(a_l))}{\overline{M}}. \tag{5}$$

When some (or all) assessments on criteria or the overall assessments become incomplete, the average similarity $\widetilde{S}(e_i)$ will become an interval with lower and upper bounds $\widetilde{S}^-(e_i)$ and $\widetilde{S}^+(e_i)$, which can be obtained from the following pair of optimization problems.

$$\text{MIN/MAX} \quad \widetilde{S}(e_i) = \frac{\sum_{l=1}^{\overline{M}} \widetilde{S}(e_i(a_l))}{\overline{M}} \tag{6}$$

$$\text{s.t.} \quad \beta_{n,i}(a_l) \le \beta^*_{n,i}(a_l) \le \beta_{n,i}(a_l) + \beta_{\Omega,i}(a_l), \tag{7}$$

$$\widetilde{\beta}_n(a_l) \le \widetilde{\beta}^*_n(a_l) \le \widetilde{\beta}_n(a_l) + \widetilde{\beta}_{\Omega}(a_l), \tag{8}$$

$$\sum_{n=1}^{N} \beta^*_{n,i} = 1, \tag{9}$$

$$\sum_{n=1}^{N} \widetilde{\beta}^*_n = 1. \tag{10}$$

In the pair of optimization problems, $\beta^*_{n,i}(a_l)$ and $\widetilde{\beta}^*_n(a_l)$ represent decision variables. The process of calculating $\widetilde{S}(e_i(a_l))$ according to Definition 2 is implicitly included in the optimization problems. It is clear that $[\widetilde{S}^-(e_i), \widetilde{S}^+(e_i)] \subseteq [0, 1]$. Without loss of generality, $[\widetilde{S}^-(e_i), \widetilde{S}^+(e_i)]$ is used for the determination of constraints on the criterion weights by following the idea that higher $[\widetilde{S}^-(e_i), \widetilde{S}^+(e_i)]$ means a larger contribution of the assessments on criterion e_i to overall assessments for all alternatives. The constraints can usually be expressed in a linear inequality way, such as bounded constraint of weights (e.g., $LB_i \le w_i \le UB_i$ $(i \in \{1, \ldots, L\})$), bounded ratio of weights (e.g., $LB_i \le w_i/w_j \le UB_i$ $(i, j \in \{1, \ldots, L\})$), and bounded difference of weights (e.g., $LB_i \le w_i - w_j \le UB_i$ $(i, j \in \{1, \ldots, L\})$) [30]. Note that the constraints learned from $[\widetilde{S}^-(e_i), \widetilde{S}^+(e_i)]$ should be consistent with the preferences of a decision maker unless such preferences are unavailable.

Within the feasible region of criterion weights satisfying the normalized constraint $\sum_{i=1}^{L} w_i = 1$ and the learned constraints, the average difference between the overall assessments provided and the aggregated assessments generated from the assessments on criteria is anticipated to be minimized. By following this principle, an optimization model is constructed to learn criterion weights.

$$\text{MIN} \quad F = \frac{\sum_{l=1}^{\overline{M}} \widetilde{D}(a_l)}{\overline{M}} \tag{11}$$

$$\text{s.t.} \quad w^* \in C(w^*), \tag{12}$$

$$0 \le w_i^* \le 1, \tag{13}$$

$$\sum_{i=1}^{L} w_i^* = 1. \tag{14}$$

In this model, $\widetilde{D}(a_l)$ represents the dissimilarity between the aggregated assessment $B(a_l)$ and the overall assessment provided $\widetilde{B}(a_l)$ calculated by using Definition 2. Hereinto, the criterion weights are involved in the process of calculating the aggregated assessment $B(a_l)$. In addition, w_i^* represents the decision variable of the criterion weight and $C(w^*) = \{w^*|A \cdot w^* \le c\}$ denotes the learned constraints on criterion weights, where A is an $R \times L$ matrix of coefficients, c is a column vector with R elements, and R is the number of constraints. The combination of assessments on criteria by using the ER rule shown in Definition 1 and criterion weights is implicitly included in the model. It is clear that F is limited to [0, 1]. Solving the model generates the optimal value F^*.

3.2. Preference learning with gold standards

What has been discussed above is on the condition that there is no gold standard for MCDM. The results of MCDM in this situation cannot be simply considered to be correct or incorrect. In some MCDM problems, however, gold standards may be available, which can be used to judge the correctness of the overall assessments and the decision results.

Gold standards represent the correct result of the decision problem. For example, pathologic findings using surgery or fine needle aspiration biopsy (FNAB) are gold standards of a radiologist diagnosing thyroid cancer. For MCDM problems with gold standards, preference learning may be different from what is presented in Section 3.1. Whether gold standards are involved in preference learning is determined by the willingness of a decision maker. In theory, gold standards can help a decision maker improve decision capability. This is the case when the decision maker wishes to analyze what might result in the difference between the overall assessments provided and gold standards. It cannot be guaranteed that each decision maker is willing to conduct such analysis.

Assumption 1. Facing MCDM problems with gold standards, a decision maker is willing to analyze the difference between his or her overall assessments and gold standards so that he or she can provide more correct assessments for future problems under the same criterion framework.

Whether a decision maker's preferences are learned from the overall assessments provided is dependent on whether Assumption 1 is satisfied. If Assumption 1 is not satisfied in consideration of the decision maker's willingness, the overall assessments and their corresponding assessments on criteria are used to learn the preferences of the decision maker through what is presented in Section 3.1.

When a decision maker wishes to accept Assumption 1, the weight of a criterion should characterize both the contribution of the assessment on the criterion to overall assessment and that to gold standard denoted by $\overrightarrow{B}(a_l) = \{(H_n, \overrightarrow{\beta}_n(a_l)), n = 1, \ldots, N; (\Omega, \overrightarrow{\beta}_\Omega(a_l))\}$. In this situation, the average similarity between $B(e_i(a_l))$ and $\widetilde{B}(a_l)$, i.e., $\widetilde{S}(e_i(a_l))$ and the one between $B(e_i(a_l))$ and $\overrightarrow{B}(a_l)$, i.e., $\overrightarrow{S}(e_i(a_l))$ are combined to form the aggregated average similarity $[\bar{S}^-(e_i), \bar{S}^+(e_i)]$. This can be obtained with the following pair of optimization problems adapted from the problems shown in Eqs. (6)–(10).

$$\text{MIN/MAX} \quad \bar{S}(e_i) = (1-\eta) \cdot \frac{\sum_{l=1}^{\overline{M}} \widetilde{S}(e_i(a_l))}{\overline{M}} + \eta \cdot \frac{\sum_{l=1}^{\overline{M}} \overrightarrow{S}(e_i(a_l))}{\overline{M}} \tag{15}$$

$$\text{s.t.} \quad \beta_{n,i}(a_l) \le \beta_{n,i}^*(a_l) \le \beta_{n,i}(a_l) + \beta_{\Omega,i}(a_l), \tag{16}$$

$$\widetilde{\beta}_n(a_l) \le \widetilde{\beta}_n^*(a_l) \le \widetilde{\beta}_n(a_l) + \widetilde{\beta}_\Omega(a_l), \tag{17}$$

$$\sum_{n=1}^{N} \beta_{n,i}^* = 1, \tag{18}$$

$$\sum_{n=1}^{N} \widetilde{\beta}_n^* = 1. \tag{19}$$

In the above problems, the parameter η represents the learning rate of the decision maker. It is also satisfied that $[\bar{S}^-(e_i), \bar{S}^+(e_i)] \subseteq [0, 1]$. The larger the value of η, the stronger the willingness of the decision maker to improve his or her decision capability with the aid of gold standards. By following this idea, the decision maker's capability to make correct or rational decisions can be used to calculate the value of η. To facilitate quantifying the capability, the reliability of the decision maker is defined as

$$R = 1 - \frac{\sum_{l=1}^{\overline{M}} \overrightarrow{\widetilde{D}}(a_l)}{\overline{M}}, \tag{20}$$

where $\overrightarrow{\widetilde{D}}(a_l)$ represents the dissimilarity between the overall assessment $\widetilde{B}(a_l)$ and gold standard $\overrightarrow{B}(a_l)$ calculated by using Definition 2. The value range of R is included in [0, 1]. There are two special situations in which the value of R is equal to 1 and 0, respectively. $R = 1$ means the decision maker is fully reliable, and $R = 0$ means the decision maker is fully unreliable. In other situations, the decision maker is partially reliable. When it is accepted that a decision maker with higher reliability is capable of making more correct decisions, the reliability of the decision maker can be regarded as an indicator of his or her learning rate, i.e.,

$$\eta = R. \tag{21}$$

There may be other indicators of the decision maker's learning rate. The learning rate shown in Eq. (21) is generated from historical overall assessments provided by the decision maker and gold standards. It can objectively reflect the willingness of the decision maker to improve his or her decision capability through gold standards. Constraints on criterion weights can then be learned from $[\bar{S}^-(e_i), \bar{S}^+(e_i)]$. Within the feasible region of criterion weights satisfying the normalized constraint $\sum_{i=1}^{L} w_i = 1$ and the learned constraints, the average difference between the overall assessments $\widetilde{B}(a_l)$ and the aggregated assessments generated from the assessments on criteria $B(a_l)$ and the average difference between gold standards $\overrightarrow{B}(a_l)$ and the aggregated assessments $B(a_l)$ are anticipated to be minimized. The two average differences are balanced by the learning rate η. The following optimization model is constructed to learn criterion weights by carrying out the minimization.

$$\text{MIN}\quad G=(1-\eta)\cdot\frac{\sum_{l=1}^{\overline{M}}\widetilde{D}(a_l)}{\overline{M}}+\eta\cdot\frac{\sum_{l=1}^{\overline{M}}\overrightarrow{D}(a_l)}{\overline{M}} \tag{22}$$

$$\text{s.t.}\quad w^*\in C(w^*), \tag{23}$$

$$0\le w_i^*\le 1, \tag{24}$$

$$\sum_{i=1}^{L} w_i^*=1. \tag{25}$$

In this model, $\overrightarrow{D}(a_l)$ represents the dissimilarity between the aggregated assessment $B(a_l)$ and gold standard $\overrightarrow{B}(a_l)$ calculated by using Definition 2. The aggregated assessment is generated by using the ER rule shown in Definition 1 and the criterion weights. Similar to the optimization model shown in Eqs. (11)–(14), G in this model is limited to [0, 1], and its optimal value is denoted by G^*.

Note that when a decision maker does not accept Assumption 1, the situation presented in Section 3.1 can be seen as a special case of what is discussed in Section 3.2. On the condition that Assumption 1 is accepted by the decision maker, the process of learning criterion weights in Section 3.1 is different from that in Section 3.2. This is because it is impossible to find a fully unreliable decision maker in practice and set his or her learning rate as 0. As a result, gold standards have a significant influence on preference learning when they are available and accepted by a decision maker.

4. Decision making through the learned preferences

The purpose of preference learning is to objectively determine criterion weights from historical decision data and then aid a decision maker in making future decisions that are consistent with his or her preferences to a maximum extent. For this purpose, this section presents how to generate a solution from the learned criterion weights.

4.1. Generation of a solution without the consideration of gold standards

To be consistent with Section 3, the situation where gold standards are not available is considered first. Solving the optimization model shown in Eqs. (11)–(14) generates F^* and a set of criterion weights. In a new MCDM problem, the set of criterion weights learned from historical decision data can be used to combine the assessments of an alternative on criteria to generate the aggregated assessment of the alternative. The ER rule shown in Definition 1 is applied in the combination process. In some situations, the aggregated assessment $B(a_l)$ is sufficient for a decision maker to make judgments on alternative a_l and provide service to the next handling of the alternative with the aid of a specific principle. For example, a radiologist can generate the diagnosis of an inpatient's thyroid nodule from the aggregated assessment through the principle of maximum belief degree. In detail, given the aggregated assessment $B(a_l) = \{(H_n, \beta_n(a_l)), n = 1, \ldots, N\}$, the principle of maximum belief degree is used to transform $B(a_l)$ into the diagnosis $\{(H_{\hat{n}}, 1)\}$ where $\hat{n} = \operatorname{argmax}\{\beta_n(a_l), n = 1, \ldots, N\}$. Such a diagnosis aids a clinician in knowing how to handle the inpatient.

Appropriate principles are not always available to aid a decision maker in generating a solution from the aggregated assessment. To address this situation, the aggregated assessment can be combined with utilities of grads $u(H_n)$ ($n = 1, \ldots, N$) to generate the minimum and maximum expected utilities of each alternative $[u^-(a_l), u^+(a_l)]$, as presented in Section 2. The remainder is how a decision maker generates a solution from the expected utilities $[u^-(a_l), u^+(a_l)]$. It is usually a problem-specific process. With the sufficient consideration of the characteristics of the decision problem, the decision maker can select to directly generate a solution from $[u^-(a_l), u^+(a_l)]$ or do it with the aid of appropriate decision rules. For example, the decision maker can generate a solution from $[u^-(a_l), u^+(a_l)]$ through the Hurwicz rule [30]. The prerequisite is that the decision maker is required to specify the optimism degree γ to transform $[u^-(a_l), u^+(a_l)]$ into the expected utility $E(a_l)$, i.e.,

$$E(a_l)=\gamma\cdot u^+(a_l)+(1-\gamma)\cdot u^-(a_l). \tag{26}$$

Here, γ is limited to [0,1].

In addition to aiding a decision maker in generating solutions to new MCDM problems, the effect of the learned criterion weights can be examined when historical data are sufficient. Suppose that a historical decision matrix with $\overline{M}$ alternatives is used to learn criterion weights and another matrix with $\overline{M}_1$ alternatives is used to examine the effect of the learned criterion weights. Under this assumption, the effect is measured by the average similarity between the aggregated assessments of the $\overline{M}_1$ alternatives derived from the learned criterion weights and the overall assessments of the $\overline{M}_1$ alternatives, which is

$$AR=1-\frac{\sum_{l=1}^{\overline{M}_1}\widetilde{D}(a_l)}{\overline{M}_1}. \tag{27}$$

Here, AR is limited to [0, 1]. AR = 1 means that the aggregated assessments generated by using the ER rule and the learned criterion weights are completely consistent with the overall assessments provided by a decision maker, while $AR = 0$ means that the aggregated assessments of all alternatives and their overall assessments are in two extreme situations. In most cases, AR is within (0, 1).

4.2. Generation of a solution with the consideration of gold standards

The above discussion is about the process of how to generate a solution when gold standards are unavailable. In the following, how to generate a solution from the learned criterion weights with the consideration of gold standards will be discussed.

The question of whether gold standards influence decisions depends on whether a decision maker accepts Assumption 1. If a decision maker makes the same decision regardless of whether he or she knows the gold standards, the optimization model shown in Eqs. (11)–(14) is solved to generate F^* and a set of criterion weights. The aggregated assessment of each alternative in a new MCDM problem is then produced by using the ER rule shown in Definition 1 and the set of learned criterion weights. In consideration of the characteristics of the decision problem, the decision maker can select to generate a solution from the aggregated assessment with the aid of specific principles. He or she can also choose to generate a solution from the expected utilities with the aid of specific decision rules, in which the expected utilities are derived from the aggregated assessments and utilities of the grades.

For a decision maker with a willingness to improve his or her decision capability derived from gold standards, the optimization model shown in Eqs. (22)–(25) is solved to generate G^* and a set of criterion weights. The aggregated assessment of each alternative in a new MCDM problem is obtained and used to generate a solution to the problem with the help of appropriate principles or decision rules, which is similar to the situation in which the decision maker does not accept Assumption 1.

Sufficient historical data allow the decision maker to examine the effect of the learned criterion weights. When the set of criterion weights is learned from the optimization model shown in Eqs. (11)–(14) or from the model shown in Eqs. (22)–(25), the examination of its effect can also be conducted by using Eq. (27), which is similar to the situation where gold standards are not available.

Sections 3 and 4 indicate that the preference learning and the decision making through the learned preferences in this study are different from those in machine learning and those in recommender systems. The preference learning in this study is not to estimate the marginal preference structures on criteria or adapt a decision maker's preference profile from his or her feedback. It is to learn the preferences of a decision maker in MCDM in the ER context when historical decision data are available. Specifically, the purpose of MCDM in this study is not only to compare alternatives but also to help decide what will be done in the next step. These analyses clarify the boundary of this study.

5. Case study

Ultrasonic examination is an important imaging technique for identifying thyroid nodules and diagnosing whether the nodules are malignant or benign lesions. As ultrasonic examination does not involve radiation, it is a popular technique for patients, and a large number of relevant data have been accumulated in clinical practices. The accumulated data reflect the expertise and experience of radiologists. Nevertheless, radiologists rarely apply the data in real diagnoses of thyroid nodules due to lack of effective ways to analyze their historical diagnostic data. They usually rely on representative cases and basic expertise to diagnose nodules. This makes it difficult for radiologists to provide consistent diagnoses, and it significantly devalues historical diagnostic data.

Clinical features (or observations) on some criteria are usually included in historical diagnostic data. This indicates that the diagnosis of thyroid nodules can be considered an MCDM problem and then the proposed method of learning preferences in MCDM in the ER context can be used to help radiologists diagnose thyroid nodules with the consideration of their expertise and experience. The detailed process is presented as follows.

5.1. Modeling of diagnosis of thyroid nodule

In clinical practice, some criteria are identified to aid radiologists in diagnosing thyroid nodules such as halo, margin, size, contour, vascularity, tallness, solid component, calcification, and echogenicity [31–34]. However, radiologists generally consider some criteria rather than all of them when diagnosing thyroid nodules. The selection of criteria is related to hospitals and their regions. In different hospitals located in different regions, different sets of criteria may be selected.

Through the expertise and experience of the third author and the analysis of historical examination reports from 2013 to 2017 in the ultrasonic department of a tertiary hospital located in Hefei, Anhui, China, five criteria are identified to help diagnose thyroid nodules. They are margin, contour, echogenicity, calcification, and vascularity, and they are denoted by e_i ($i = 1, \ldots, 5$). In the process of identifying the five criteria, the third author also communicated with the representative radiologists in the department to make the criteria commonly accepted. A radiologist observes the nodule of a patient on the five criteria to comprehensively provide the extent to which he or she has a suspicion that the nodule is a malignant lesion. It is difficult for a radiologist to provide a precise possibility that the identified nodule is a malignant lesion. To facilitate the diagnosis of a thyroid nodule, the imaging reporting and data system (TIRADS) [34–37] has been developed based on Horvath et al.'s TIRADS. The relevant details about the TIRADS are presented in Table B.1 of Appendix B of supplementary material.

The TIRADS shown in Table B.1 builds a bridge between radiologists and clinicians. With the aid of the TIRADS shown in Table B.1, a radiologist provides any of the TIRADS categories as his or her overall assessment, i.e., the overall diagnosis of the nodule identified for a patient. A clinician understands the overall diagnosis provided by the radiologist through the findings, cancer risk, and recommendations shown in Table B.1. The description of clinical features (or observations) associated with the identified thyroid nodule in an examination report reflects the assessments for the nodule, i.e., the diagnoses of the nodule on the five criteria. Facing an examination report, a clinician does not decide what will be done for the patient associated with the report simply from the overall diagnosis provided by a radiologist. The clinician conducts it by considering the overall diagnosis and the diagnoses on the five criteria. The clinician proceeds in this way because he or she aims to avoid potential incorrect operations caused by the inconsistency between the aggregation of the diagnoses on the five criteria and the overall diagnosis. In the hospital, clinicians have generally collaborated with radiologists for a long time, usually more than 2-to-5 years, and thus they are familiar with the radiologists. Under such conditions, when facing examination reports with similar diagnoses on the five criteria and different overall diagnoses provided by a radiologist, a clinician is not willing to believe the overall diagnoses in the reports and has difficulties in handling the patient with the identified nodule. Meanwhile, a radiologist may provide different overall diagnoses when facing similar nodules. It is difficult for the radiologist to recall similar cases and their corresponding diagnoses to provide the overall diagnosis to the nodule observed currently within a limited time. This gives rise to an important issue about how to improve the consistency between the aggregation of the assessments transformed from observations on the five criteria and the overall assessment provided.

To address this issue, the proposed method of learning preferences in MCDM in the ER context presented in Section 3 is used to learn the criterion weights of the radiologists in the multi-criteria diagnosis of thyroid nodules. For this purpose, both the assessment on each of the five criteria and the overall assessment are expressed as belief distributions on the set of grades $\Omega = \{H_1, H_2, \ldots, H_8\}$ = {TIRADS 3, TIRADS 4A-1, TIRADS 4A-2, TIRADS 4B-1, TIRADS 4B-2, TIRADS 4B-3, TIRADS 4C, TIRADS 5}. As mentioned above, TIRADS categories are used to characterize the overall assessments, thus they can be represented by $\{(H_n, 1)\}$ ($n \in \{1, \ldots, 8\}$). However, the assessments on the five criteria are not provided by the radiologists. They can be transformed from the observations recorded in the examination reports.

Guided by the expertise and experience of the third author and his communication with the representative radiologists in the department, effective clinical features associated with nodules are identified from historical examination reports in the period from 2013 to 2017. The relationships between the clinical features of a nodule and TIRADS categories are also constructed in a similar way to transform the observations of the nodule into its assessments. To save space, the constructed relationships are not presented. Through the relationships, the assessments on criteria can also be represented by $\{(H_n, 1)\}$. Note that the reason why the assessments are not represented by general belief distributions with nonzero belief degrees on multiple grades is that the radiologists cannot accept it. As shown in Table B.1, the cancer risk of each TIRADS category is an interval instead of a precise number. If the radiologists use general belief distributions to describe the overall assessments of nodules, the clinicians will find it very difficult to understand the assessments and judge the cancer risk of the nodules, and thus they will be unsure about what should be done for the patients with the nodules.

5.2. Learning of the preferences of radiologists

On the condition that the assessments of the thyroid nodules on the five criteria and the corresponding overall assessments are obtained from historical examination reports, it is feasible to learn the criterion weights of the radiologists in terms of what is presented in Section 3.

In this study, the examination reports of three radiologists in the department from 2013 to 2017 are collected, in which the description of clinical features of inpatients' nodules, the overall assessments of the three radiologists, and the corresponding pathologic findings are included. Pathologic findings are regarded as the gold standards for the diagnosis of thyroid nodules, which are generated when inpatients undergo FNAB. Under the conditions, what is presented in Section 3.2 is used to learn the preferences of the three radiologists and the effect of the learned preferences is examined.

The policies of the hospital encourage the radiologists to monitor and improve their own diagnostic performance. As a result, the three radiologists are willing to analyze the difference between the overall assessments and the pathologic findings to improve their diagnostic capability. Assumption 1 is satisfied and the pathologic findings are involved in learning the preferences of the three radiologists. To facilitate the preference learning, the malignant and benign findings confirmed by FNAB are represented by $\{(H_8, 1)\}$ and $\{(H_1, 1)\}$, respectively. Meanwhile, through the expertise and experience of the third author and his communication with the three radiologists, the probability assignment approach [38] is used to set the utility of each grade H_n, which is $u(H_n)$ $(n = 1, \ldots, 8) = (0, 1/7, 2/7, 3/7, 4/7, 5/7, 6/7, 1)$.

Suppose that the three radiologists are represented by D_j $(j = 1, 2, 3)$, their examination reports from 2013 to 2016 are used to learn their criterion weights, and their examination reports in 2017 are used to examine the effect of their learned criterion weights. The number of diagnostic records of thyroid nodules provided by the three radiologists in each of the five years is presented in Table B.2. For each radiologist, only 5 samples of the diagnostic records of thyroid nodules and the corresponding pathologic findings are presented as examples to save space, as shown in Table B.3.

To learn the criterion weights of the three radiologists from their diagnostic records in the period from 2013 to 2016, their learning rates are first calculated by using Eqs. (20)–(21), which are presented in Table 1. It can be known from Section 5.1 that the assessments on the five criteria and the overall assessments provided are complete. This indicates that $\bar{S}(e_i)$ $(i = 1, \ldots, 5)$ can be directly obtained from Eq. (15) rather than by solving the optimization model shown in Eqs. (15)–(19). To facilitate the determination of the criterion weights, $\bar{S}(e_i)$ is normalized to be $\bar{\bar{S}}(e_i)$. Considering $\bar{\bar{S}}(e_i)$, the three radiologists offer the constraints on the five criteria through their expertise and the help of the third author. The constraints and the relevant $\bar{S}(e_i)$ and $\bar{\bar{S}}(e_i)$ for the three radiologists are presented in Table 1.

The obtained constraints on the criterion weights from the three radiologists can be incorporated into the optimization model shown in Eqs. (22)–(25) to learn the criterion weights. This is feasible in theory but irrational in practice. As presented in Section 5.1, the overall assessments are provided by the radiologists as $\{(H_n, 1)\}$. The aggregated assessments are used as recommendations to aid the radiologists in diagnosing thyroid nodules, thus, they also should be expressed as $\{(H_n, 1)\}$. For this purpose, after a discussion of the third author with the three radiologists, the principle of maximum belief degree is adopted to transform the aggregated assessments into those expressed as $\{(H_n, 1)\}$. The details can be found in Section 4.1. When the principle of maximum belief degree is followed by the optimization model shown in Eqs. (22)–(25), solving the model generates G^* and the corresponding set of criterion weights, as presented in Table 2.

To examine the effect of their learned criterion weights, the weights are used to combine the assessments on the five criteria derived from the examination reports in 2017. By following the principle of maximum belief degree, the aggregated assessments of thyroid nodules are transformed into those expressed as $\{(H_n, 1)\}$. From the transformed aggregated assessments and the overall assessments provided by the three radiologists, the *AR* of the three radiologists is calculated by using Eq. (27) and presented in Table 4. Suppose that the *AR* of the three radiologists is denoted by AR_j $(j = 1, 2, 3)$. It is easily found that $AR_2 > AR_1 > AR_3$. Meanwhile, suppose that the learning rates of the three radiologists are denoted by η_j $(j = 1, 2, 3)$. Then it can be known from Table 2 that $\eta_2 > \eta_1 > \eta_3$. This seems to mean that the stronger the capability of a radiologist to learn from pathologic findings, the better the effect of learning criterion weights from the examination reports of the radiologist in the period from 2013 to 2016. In other words, the transformed aggregated assessments generated using the learned criterion weights and the principle of maximum belief degree may be preferable to reflect the real assessments of a radiologist with a high learning rate in the period from 2013 to 2016.

Note that the transformed aggregated assessments are only considered as recommendations to aid the radiologists in diagnosing thyroid nodules rather than to replace radiologists. The final overall assessments are provided by the radiologists in accordance with their expertise and experience. The benefits of learning the criterion weights from the historical diagnoses of the radiologists and generating diagnostic recommendations are to help the radiologists offer the overall diagnoses consistent with their historical diagnoses and to effectively avoid incorrect or irrational diagnoses that are clearly inconsistent with their expertise and experience. Specifically, when a radiologist faces a new case that he or she has rarely or never encountered, the recommendations derived from his or her historical examination reports may mislead the radiologist and be meaningless. In general, more historical examination reports could generate recommendations that are more satisfactory to radiologists because a large number of different types of cases are covered by the reports.

5.3. Partition of historical examination reports

Table 2 shows that learning the criterion weights of the three radiologists is not conducted on a large scale. If a large number of historical examination reports are accumulated and collected to learn the criterion weights of a radiologist, directly solving the optimization model shown in Eqs. (22)–(25) may be very time-consuming. A feasible approach in this situation is to split historical examination reports into several parts and learn the set of criterion weights from each part of the reports through the model. Multiple sets of criterion weights are then combined for application in generating recommendations for the radiologists. To demonstrate the idea of separation-integration, the examination reports of the three radiologists in the period from 2013 to 2016 are split into two parts, i.e., the reports in the period from 2013 to 2014 and those in the period from 2015 to 2016.

By respectively applying the reports in the period from 2013 to 2014 and those in the period from 2015 to 2016 to learn the criterion weights of the three radiologists, relevant results are obtained and presented in Tables 3 and 4. For simplicity, the constraints on the criterion weights of the three radiologists are omitted. The effects of the criterion weights of the three radiologists learned from their reports in the period from 2013 to 2014 and from those in the period from 2015 to 2016 are

Table 1
The constraints on the criterion weights of the three radiologists and the relevant intermediate results.

Radiologists	η	$\bar{S}(e_i)$	$\bar{\bar{S}}(e_i)$	$C(w)$
D_1	0.7888	(0.7053, 0.8036, 0.6503, 0.7415, 0.5422)	(0.2049, 0.2334, 0.1889, 0.2154, 0.1575)	$\{w_2 \le 2\cdot w_5, w_5 \le w_3, w_3 \le w_1, w_1 \le w_4, w_4 \le w_2\}$
D_2	0.8275	(0.7611, 0.7793, 0.5470, 0.7278, 0.6120)	(0.2221, 0.2274, 0.1596, 0.2124, 0.1786)	$\{w_2 \le 2\cdot w_3, w_3 \le w_5, w_5 \le w_4, w_4 \le w_1, w_1 \le w_2\}$
D_3	0.7507	(0.7714, 0.7877, 0.6772, 0.6999, 0.6006)	(0.2181, 0.2227, 0.1915, 0.1979, 0.1698)	$\{w_2 \le 2\cdot w_5, w_5 \le w_3, w_3 \le w_4, w_4 \le w_1, w_1 \le w_2\}$

Table 2
The learned criterion weights of the three radiologists and the relevant examination results from their examination reports in the period from 2013 to 2016.

Radiologists	G^*	Learned criterion weights	AR
D_1	0.1944	(0.1667, 0.3332, 0.1667, 0.1667, 0.1667)	0.8273
D_2	0.1954	(0.2848, 0.2856, 0.1428, 0.144, 0.1428)	0.8619
D_3	0.1917	(0.2012, 0.2643, 0.20115, 0.2012, 0.132152)	0.8124

Table 3
The learned criterion weights of the three radiologists and the relevant results from their examination reports in the period from 2013 to 2014.

Radiologists	η	G^*	Learned criterion weights	AR
D_1	0.7873	0.1774	(0.2104, 0.2458, 0.2104, 0.2104, 0.123)	0.7965
D_2	0.7996	0.2106	(0.2134, 0.2698, 0.1517, 0.2134, 0.1517)	0.8595
D_3	0.7086	0.1911	(0.205, 0.2466, 0.205, 0.205, 0.1384)	0.8124

Table 4
The learned criterion weights of the three radiologists and the relevant results from their examination reports in the period from 2015 to 2016.

Radiologists	η	G^*	Learned criterion weights	AR
D_1	0.7904	0.2107	(0.17135, 0.3146, 0.17135, 0.17135, 0.17135)	0.8273
D_2	0.847	0.1797	(0.2806, 0.2832, 0.1442, 0.1478, 0.1442)	0.8619
D_3	0.7808	0.2058	(0.25, 0.25, 0.125, 0.25, 0.125)	0.7547

examined, and the relevant results are also presented in Tables 3 and 4. It can be found from Table 3 that $\eta_2 > \eta_1 > \eta_3$ and $AR_2 > AR_1 > AR_3$ hold simultaneously. However, Table 4 shows that $\eta_2 > \eta_1 > \eta_3$ and $AR_2 > AR_3 > AR_1$. The findings shown in Tables 3 and 4 mean that the relationship between η and AR is associated with data collected in a specific period.

Suppose that the learned criterion weights of the three radiologists in Table 3 are denoted by w_j^{34} ($j = 1, 2, 3$) and those in Table 6 by w_j^{56} ($j = 1, 2, 3$). Assume that the numbers of diagnostic records of thyroid nodules provided by the three radiologists in the period from 2013 to 2014 are denoted by N_j^{34} ($j = 1, 2, 3$) and those in the period from 2015 to 2016 by N_j^{56} ($j = 1, 2, 3$). The aggregated criterion weights of the three radiologists are then calculated as

$$w_{j,i}^{36} = w_{j,i}^{34}\cdot\frac{N_j^{34}}{N_j^{34}+N_j^{56}} + w_{j,i}^{56}\cdot\frac{N_j^{56}}{N_j^{34}+N_j^{56}}, j = 1, 2, 3, i = 1, \ldots, 5. \tag{28}$$

Note that the aggregation of w_j^{34} and w_j^{56} is under the assumption that the contributions of w_j^{34} and w_j^{56} to $w_{j,i}^{36}$ are proportional to N_j^{34} and N_j^{56}. When other assumptions are accepted by the three radiologists, other ways will be developed to aggregate w_j^{34} and w_j^{56}. The effects of the aggregated criterion weights of the three radiologists are examined by using their examination reports in 2017. The relevant results are shown in Table 5.

Making a comparison between Tables 2 and 5, it is found that AR_1 and AR_2 decrease but AR_3 increases after the process of separation-integration. This highlights the influence of the process of separation-integration on the value of AR.

An interesting issue is whether different divisions of historical examination reports result in different aggregated criterion weights and different effects of the weights. To address this issue, the examination reports of radiologist D_1 in the period from 2013 to 2016 are split into the reports in the period from 2013 to 2015 and those in 2016. The reports in these two parts are used to learn the criterion weights of radiologist D_1, and the effects of the learned weights are examined by using the reports of D_1 in 2017. The relevant results are shown in Table 6.

Table 5
The aggregated criterion weights of the three radiologists and the relevant examination results.

Radiologists	Aggregated criterion weights	AR
D_1	(0.1917, 0.2787, 0.1917, 0.1917, 0.1462)	0.8105
D_2	(0.2529, 0.2777, 0.1473, 0.1748, 0.1473)	0.8499
D_3	(0.2312, 0.2486, 0.1584, 0.2312, 0.1306)	0.8124

Through Eq. (28), the aggregated criterion weights of radiologist D_1 is obtained as (0.1665, 0.3327, 0.1665, 0.1678, 0.1665), whose effect is examined to generate $AR_1 = 0.8273$. The aggregated criterion weights derived from the division of the examination reports in 2013–2015 and 2016 are different from those shown in Table 5 and very close to the learned criterion weights shown in Table 2. This explains why AR_1 is different from that shown in Table 6 and the same as that shown in Table 2. These analyses indicate that different annual divisions of examination reports result in different criterion weights and different effects of the weights. The question of how to split examination reports is dependent on their time characteristic. By analyzing the other characteristics of the examination reports, other ways to split examination reports may be found. Which division is better is closely associated with the expertise and experience of the radiologists because the purpose of splitting the examination reports is to learn the preferences of the radiologists.

Another interesting issue is how to split examination reports when there is no clear characteristic or no characteristic that is accepted by the radiologists. Randomly splitting the examination reports may be a feasible choice in this situation. Similar to the division by considering the time characteristic, different random

Table 6
The learned criterion weights of radiologist D_1 and the relevant results.

Division of reports	η	G^*	Learned criterion weights	AR
2013–2015	0.7864	0.1791	(0.1667, 0.3332, 0.1667, 0.1667, 0.1667)	0.8273
2016	0.793	0.2211	(0.1661, 0.332, 0.1661, 0.1697, 0.1661)	0.8273

Table 7
The learned criterion weights of radiologist D_1 and the relevant results from the random divisions of the radiologist's diagnostic records in the period from 2013 to 2016.

No	Records	η	G^*	Learned criterion weights	AR
1	172	0.8056	0.0983	(0.1429, 0.2857, 0.14285, 0.2857, 0.14285)	0.7974
1	177	0.7724	0.1014	(0.1758, 0.3242, 0.1621, 0.1758, 0.1621)	0.8273
2	193	0.7728	0.1105	(0.1667, 0.3332, 0.1667, 0.1667, 0.1667)	0.8273
2	156	0.8086	0.084	(0.162, 0.3239, 0.162, 0.1901, 0.162)	0.8273
3	183	0.7736	0.1028	(0.1667, 0.3332, 0.1667, 0.1667, 0.1667)	0.8273
3	166	0.8055	0.0851	(0.199, 0.2825, 0.1586, 0.2057, 0.1542)	0.8105
4	169	0.7929	0.0912	(0.1667, 0.3332, 0.1667, 0.1667, 0.1667)	0.8273
4	180	0.7849	0.1071	(0.1454, 0.2819, 0.1454, 0.2819, 0.1454)	0.7955
5	174	0.812	0.0887	(0.1864, 0.2204, 0.1864, 0.2204, 0.1864)	0.774
5	175	0.7657	0.1054	(0.1667, 0.3332, 0.1667, 0.1667, 0.1667)	0.8273

Table 8
The aggregated criterion weights of radiologist D_1 and the relevant results from the random divisions of the radiologist's diagnostic records in the period from 2013 to 2016.

No	Aggregated criterion weights	AR
1	(0.1596, 0.3052, 0.1526, 0.23, 0.1526)	0.8161
2	(0.1646, 0.329, 0.1646, 0.1772, 0.1646)	0.8273
3	(0.1821, 0.3091, 0.1628, 0.1852, 0.1608)	0.8105
4	(0.1557, 0.3068, 0.1557, 0.2261, 0.1557)	0.8161
5	(0.1765, 0.277, 0.1765, 0.1935, 0.1765)	0.8105

divisions of the examination reports may result in different aggregated criterion weights and different effects of the weights. To verify this, it is necessary to redo the experiments by randomly splitting the examination reports of radiologist D_1 in the period from 2013 to 2016 into two parts five times. The learned criterion weights and their effects are shown in Table 7, in which the relevant results are also presented. Table 7 shows that the criterion weights learned from one part of each division are different from those from the other part; however, the weights learned from some parts of different divisions may be equal to one another. More interestingly, the same criterion weights result in the same effect although different weights may also generate the same effect.

Through Eq. (28), the criterion weights learned from the two parts of each division are combined to generate the aggregated criterion weights of each division. The weights and their effects are presented in Table 8. By observing Tables 7 and 8, it can be found that the effect of the aggregated criterion weights falls in between the effects of two sets of weights learned from the two parts in a random division. For the choice of random divisions, larger effect of the aggregated criterion weights is preferred. More importantly, the size of the reports for learning criterion weights determines how many parts into which the reports should be divided. The learning process must be completed within a limited or acceptable period. Otherwise, random division will not be very meaningful. All these are problem-specific.

Observing Tables 2 and 8, it is found that the effects of the aggregated criterion weights shown in Table 8 are less than the effect of radiologist D_1 shown in Table 2. A hypothesis arises from this observation, which is the effect of the aggregated criterion weights derived from a random division is always less than or equal to the effect of each of the three radiologists shown in Table 2. As mentioned above, large effect of the aggregated criterion weights derived from a random division is preferred. Following this principle, it can be derived that the largest one among the

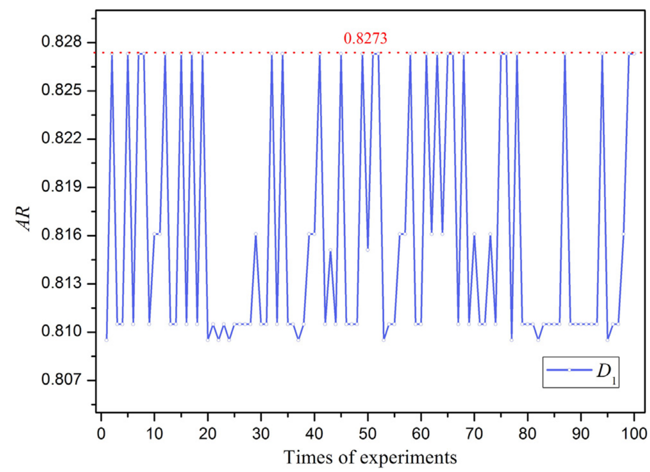

Fig. 1. The experimental effects of radiologist D_1 from 100 times of random divisions.

effects of the aggregated criterion weights derived from random divisions is the most desirable and may be equal to the effect of each of the three radiologists shown in Table 2 if the hypothesis is confirmed. Because the hypothesis is associated with the examination reports of the three radiologists in the period from 2013 to 2016, theoretical confirmation may be difficult. In this situation, simulation experiments are conducted to confirm the hypothesis. For each of the three radiologists, 100 times of random divisions are performed to generate the aggregated criterion weights and the corresponding effects. The experimental effects of the three radiologists are plotted in Figs. 1–3, respectively.

Figs. 1–3 show that the experimental effects of the three radiologists are always less than or equal to the effects of the three radiologists shown in Table 2. Meanwhile, the effects of the three radiologists shown in Table 2 are reached many times in simulation experiments. The findings confirm the hypothesis. In general, simulation based on random divisions offers an effective way to estimate the effect of a radiologist when it is difficult to estimate the effect based on the collected data of the radiologist in a specific period.

6. Conclusions

To aid a decision maker in providing consistent overall assessments, this paper proposes a method of learning the preferences

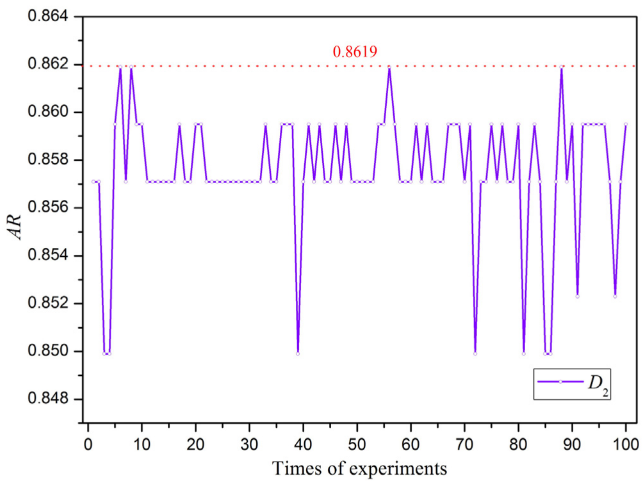

Fig. 2. The experimental effects of radiologist D_2 from 100 times of random divisions.

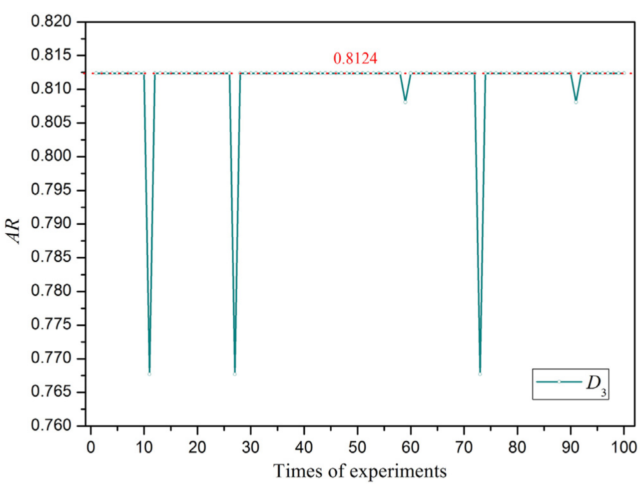

Fig. 3. The experimental effects of radiologist D_3 from 100 times of random divisions.

of the decision maker from the observations on each criterion and the overall assessments provided. On the condition that the assessments on criteria can be transformed from the corresponding observations, how to learn the constraints on criterion weights from the assessments on criteria and the overall assessments and how to learn criterion weights within the feasible region formed by the learned constraints are presented. In another situation in which gold standards are available, such as pathologic findings for the diagnoses of thyroid nodules, the learning of the constraints on criterion weights and the learning of criterion weights are also presented. The generation of solutions from the learned criterion weights and the examination of the effect of the learned weights with and without the consideration of gold standards are introduced. A case study is conducted to demonstrate the application of preference learning in MCDM to the diagnosis of thyroid nodules in the ultrasonic department of a tertiary hospital located in Hefei, Anhui, China.

The above analysis of the main work indicates that this study focuses on the preference learning in MCDM in the ER context and the application of MCDM to both comparing alternatives and deciding what will be done in the next step when historical decision data are available. This differentiates this study from existing studies on preference learning and MCDM.

Based on the above analysis of the main work, the implications of this study are summarized as follows. (1) Data can reflect the expertise and experience of a decision maker, which are characterized by the preferences of the decision maker. The accumulation of data is helpful to learn the preferences of the decision maker to make decisions consistent with his or her expertise and experience. (2) When there exist gold standards for a specific decision problem, which indicate the correct result of the decision problem, the gold standards corresponding to the historical decision data of a decision maker should be collected and used to help improve the decision maker's capability to make correct or rational decisions. (3) Large volumes of data may make preference learning uneasy. The regular separation-integration strategy is encouraged to be adopted to conduct preference learning if the characteristics of the data associated with data separation are available and accepted by the decision maker. Otherwise, the random separation-integration strategy is encouraged to be adopted when there are sufficient computational resources and relatively loose constraints on computation time.

In the future, the work in this paper will be extended from the theoretical and applicable perspectives. In theory, preference learning in group decision making in the ER context will be investigated when gold standards are available or unavailable. In application, the proposed method of learning preferences in MCDM in the ER context will be used to help generate the consistent diagnoses of other diseases, such as breast cancer, liver cancer and lung cancer. To facilitate the application, large volumes of diagnostic data will be collected and then the extraction of the disease features and the transformation of the features into the assessments on criteria will be investigated.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research is supported by the National Natural Science Foundation of China (Grant Nos. 71622003, 71571060, and 72001063). In addition, the last two authors are grateful for the support by NSFC, China-Zhejiang Joint Fund for the Integration of Industrialization and Informatization (U1709215).

Appendix A. Supplementary concepts and data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.asoc.2021.107109.

References

[1] S. Corrente, S. Greco, M. Kadziński, R. Słowiński, Robust ordinal regression in preference learning and ranking, Mach. Learn. 93 (2–3) (2013) 381–422, http://dx.doi.org/10.1007/s10994-013-5365-4.

[2] H.Q. Lu, C.C. Chen, M. Kong, H.Y. Zhang, Z. Zhao, Social recommendation via multi-view user preference learning, Neurocomputing 216 (2016) 61–71, http://dx.doi.org/10.1016/j.neucom.2016.07.011.

[3] L. Marin, A. Moreno, D. Isern, Automatic preference learning on numeric and multi-valued categorical attributes, Knowl.-Based Syst. 56 (2014) 201–215, http://dx.doi.org/10.1016/j.knosys.2013.11.012.

[4] B. Dutta, S.D. Dao, L. Martínez, M. Goh, An evolutionary strategic weight manipulation approach for multi-attribute decision making: TOPSIS method, Internat. J. Approx. Reason. 129 (2021) 64–83, http://dx.doi.org/10.1016/j.ijar.2020.11.004.

[5] S.Q. Wu, M. Wu, Y.C. Dong, H.M. Liang, S.H. Zhao, The 2-rank additive model with axiomatic design in multiple attribute decision making, European J. Oper. Res. 287 (2) (2020) 536–545, https://doi.org/10.1016/j.ejor.2020.04.011.

[6] Y. Ataei, A. Mahmoudi, M.R. Feylizadeh, D.F. Li, Ordinal priority approach (OPA) in multiple attribute decision-making, Appl. Soft Comput. J. 86 (2020) 105893, https://doi.org/10.1016/j.asoc.2019.105893.
[7] N. Ploskas, J. Papathanasiou, A decision support system for multiple criteria alternative ranking using TOPSIS and VIKOR in fuzzy and nonfuzzy environments, Fuzzy Sets and Systems 377 (2019) 1–30, https://doi.org/10.1016/j.fss.2019.01.012.
[8] S. Corrente, S. Greco, R. Słowiński, Handling imprecise evaluations in multiple criteria decision aiding and robust ordinal regression by n-point intervals, Fuzzy Optim. Decis. Mak. 16 (2017) 127–157, http://dx.doi.org/10.1007/s10700-016-9244-x.
[9] M. Hasan, I.E. Büyüktahtakin, E. Elamin, A multi-criteria ranking algorithm (MCRA) for determining breast cancer therapy, Omega 82 (2019) 83–101, http://dx.doi.org/10.1016/j.omega.2017.12.005.
[10] J.C. Butler, D.J. Morrice, P.W. Mullarkey, A multiple attribute utility theory approach to ranking and selection, Manage. Sci. 47 (6) (2001) 800–816, http://dx.doi.org/10.1287/mnsc.47.6.800.9812.
[11] R.L. Keeney, Common mistakes in making value trade-offs, Oper. Res. 50 (6) (2002) 935–945, http://dx.doi.org/10.1287/opre.50.6.935.357.
[12] J.B. Lan, H.Y. Zou, M.M. Hu, Dominance degrees for intervals and their application in multiple attribute decision-making, Fuzzy Sets and Systems 383 (2020) 146–164, http://dx.doi.org/10.1016/j.fss.2019.07.001.
[13] A.F. Tehrani, W.W. Cheng, E. Hüllermeier, Preference learning using the choquet integral: the case of multipartite ranking, IEEE Trans. Fuzzy Syst. 20 (6) (2012) 1102–1113, http://dx.doi.org/10.1109/TFUZZ.2012.2196050.
[14] G.N. Yannakakis, M. Maragoudakis, J. Hallam, Preference learning for cognitive modeling: a case study on entertainment preferences, IEEE Trans. Syst. Man Cybern. A 39 (6) (2009) 1165–1175, http://dx.doi.org/10.1109/TSMCA.2009.2028152.
[15] J. Lu, D. Wu, M. Mao, W. Wang, G. Zhang, Recommender system application developments: a survey, Decis. Support Syst. 74 (2015) 12–32, http://dx.doi.org/10.1016/j.dss.2015.03.008.
[16] X. Zhou, G.B. Guo, S. Zhu, L. Yuan, Multi-facet user preference learning for fine-grained item recommendation, Neurocomputing 385 (2020) 258–268, http://dx.doi.org/10.1016/j.neucom.2019.12.089.
[17] Z.Y. Zhang, Y. Liu, G.D. Xu, H.Q. Chen, A weighted adaptation method on learning user preference profile, Knowl.-Based Syst. 112 (2016) 114–126, http://dx.doi.org/10.1016/j.knosys.2016.09.003.
[18] Z. Zhao, H.Q. Lu, D. Cai, X.F. He, Y.T. Zhuang, User preference learning for online social recommendation, IEEE Trans. Knowl. Data Eng. 28 (9) (2016) 2522–2534, http://dx.doi.org/10.1109/TKDE.2016.2569096.
[19] M. Aggarwal, On learning of weights through preferences, Inform. Sci. 321 (2015) 90–102, http://dx.doi.org/10.1016/j.ins.2015.05.034.
[20] A. Abdelli, L. Mokdad, Y. Hammal, Dealing with value constraints in decision making using MCDM methods, J. Comput. Sci. 44 (2020) 101154, http://dx.doi.org/10.1016/j.jocs.2020.101154.
[21] V. Campos-Guzmán, M.S. García-Cáscales, N. Espinosa, A. Urbina, Life cycle analysis with multi-criteria decision making: A review of approaches for the sustainability evaluation of renewable energy technologies, Renew. Sustain. Energy Rev. 104 (2019) 343–366, http://dx.doi.org/10.1016/j.rser.2019.01.031.
[22] M. Chalgham, I. Khatrouch, M. Masmoudi, O.C. Walha, A. Dammak, Inpatient admission management using multiple criteria decision-making methods, Oper. Res. Health Care 23 (2019) 100173, http://dx.doi.org/10.1016/j.orhc.2018.10.001.
[23] T. Li, A. Li, X.P. Guo, The sustainable development-oriented development and utilization of renewable energy industry-A comprehensive analysis of MCDM methods, Energy 212 (2020) 118694, http://dx.doi.org/10.1016/j.energy.2020.118694.
[24] J.B. Yang, Y.M. Wang, D.L. Xu, K.S. Chin, The evidential reasoning approach for MADA under both probabilistic and fuzzy uncertainties, European J. Oper. Res. 171 (1) (2006) 309–343, http://dx.doi.org/10.1016/j.ejor.2004.09.017.
[25] C. Fu, W.J. Chang, D.L. Xu, S.L. Yang, An evidential reasoning approach based on criterion reliability and solution reliability, Comput. Ind. Eng. 128 (2019) 401–417, https://doi.org/10.1016/j.cie.2018.12.036.
[26] C. Fu, W.J. Chang, M. Xue, S.L. Yang, Multiple criteria group decision making with belief distributions and distributed preference relations, European J. Oper. Res. 273 (2019) 623–633, https://doi.org/10.1016/j.ejor.2018.08.012.
[27] J.B. Yang, D.L. Xu, Evidential reasoning rule for evidence combination, Artificial Intelligence 205 (2013) 1–29, http://dx.doi.org/10.1016/j.artint.2013.09.003.
[28] D.L. Xu, An introduction and survey of the evidential reasoning approach for multiple criteria decision analysis, Ann. Oper. Res. 195 (1) (2012) 163–187, http://dx.doi.org/10.1007/s10479-011-0945-9.
[29] C. Fu, J.B. Yang, S.L. Yang, A group evidential reasoning approach based on expert reliability, European J. Oper. Res. 246 (3) (2015) 886–893, http://dx.doi.org/10.1016/j.ejor.2015.05.042.
[30] Z.Z. Jiang, R.Y. Zhang, Z.P. Fan, X.H. Chen, A fuzzy matching model with hurwicz criteria for one-shot multi-attribute exchanges in E-brokerage, Fuzzy Optim. Decis. Mak. 14 (2015) 77–96, http://dx.doi.org/10.1007/s10700-014-9189-x.
[31] C. Cappelli, M. Castellano, I. Pirola, D. Cumetti, B. Agosti, E. Gandossi, The predictive value of ultrasound findings in the management of thyroid nodules, Qjm Mon. J. Assoc. Physicians 100 (1) (2007) 29–35, http://dx.doi.org/10.1093/qjmed/hcl121.
[32] B.K. Chan, T.S. Desser, I.R. McDougall, R.J. Weigel, R.B. Jeffrey, Common and uncommon sonographic features of papillary thyroid carcinoma, J. Ultrasound Med. 22 (10) (2003) 1083–1090, http://dx.doi.org/10.7863/jum.2003.22.10.1083.
[33] M.C. Frates, C.B. Benson, J.W. Charboneau, E.S. Cibas, O.H. Clark, B.G. Coleman, Management of thyroid nodules detected at us: society of radiologists in ultrasound consensus conference statement, Radiology 237 (3) (2005) 794–800, http://dx.doi.org/10.1148/radiol.2373050220.
[34] J.Y. Kwak, K.H. Han, J.H. Yoon, H.J. Moon, E.J. Son, S.H. Park, Thyroid imaging reporting and data system for us features of nodules: a step in establishing better stratification of cancer risk, Radiology 260 (3) (2011) 892–899, http://dx.doi.org/10.1148/radiol.11110206.
[35] E. Horvath, S. Majlis, R. Rossi, C. Franco, J.P. Niedmann, A. Castro, An ultrasonogram reporting system for thyroid nodules stratifying cancer risk for clinical management, J. Clin. Endocrinol. Metabol. 94 (5) (2009) 1748–1751, http://dx.doi.org/10.1210/jc.2008-1724.
[36] E. Horvath, C.F. Silva, S. Majlis, I. Rodriguez, V. Skoknic, A. Castro, Prospective validation of the ultrasound based TIRADS (thyroid imaging reporting and data system) classification: results in surgically resected thyroid nodules, Eur. Radiol. 27 (6) (2017) 2619–2628, http://dx.doi.org/10.1007/s00330-016-4605-y.
[37] J.Y. Park, H.J. Lee, H.W. Jang, H.K. Kim, J.H. Yi, W. Lee, A proposal for a thyroid imaging reporting and data system for ultrasound features of thyroid carcinoma, Thyroid 19 (11) (2009) 1257–1264, http://dx.doi.org/10.1089/thy.2008.0021.
[38] W. Winston, Operations Research: Applications and Algorithms, Tsinghua University Press, Beijing, 2011.