ID-GMLM: Intelligent Decision-Making with Integrated Graph Models and Large Language Models

Zhenhua Meng, Fanshen Meng, Rongheng Lin*, Budan Wu

State Key Laboratory of Networking and Switching Technology
Beijing University of Posts and Telecommunications
{zhmeng, mengfanshen, rhlin, wubudan}@bupt.edu.cn

Abstract

Multi-criteria decision making (MCDM) and preference learning (PL) are crucial subfields of intelligent decision-making, both aiming to aid decision-makers (DMs) in selecting, classifying, or ranking alternatives. While MCDM and PL can complement each other to some extent, existing approaches combining MCDM and PL often struggle with large data volumes and complex relational information. To address this, we propose a novel approach called **ID-GMLM** that integrates graph models and large language models (LLMs) for intelligent decision-making. It reformulates decision-making as a high-parallelism ranking function in the graph domain, using graph neural networks (GNNs) to learn and understand complex relationships between alternatives or criteria, and LLMs to parse and quantify the preferences of DMs. ID-GMLM features a multi-task learning framework that optimizes the primary task of predicting alternative rankings while modeling criterion interactions through the auxiliary task. Additionally, ID-GMLM incorporates a parameter tuning network based on criterion weights and an attention network, allowing the model to adaptively adjust to the context of the current task and the evolving preferences of DMs. Experiments on benchmark datasets demonstrate that ID-GMLM achieves significant performance improvements, inheriting the interpretability and intuitive appeal of MCDM while leveraging the computational efficiency and high accuracy of PL.

Introduction

In today's information age, intelligent decision support systems have emerged as crucial tools for aiding complex decision-making processes (Martyn and Kadziński 2023). Theses systems help people make better decisions through data analysis and model prediction, particularly in fields such as finance, healthcare, and transportation, demonstrating extensive application potential (Fu et al. 2020; Simeone, Zeng, and Caggiano 2021). With the rapid development of big data and artificial intelligence technologies, intelligent decision-making is increasingly focused on integrating the computational capabilities of machines with the intuitive judgment of human decision-makers (DMs). This integration improves decision efficiency through continuous learning and optimization. Moreover, as explainable machine learning continues to advance, enhancing the transparency of models becomes essential for the practical application of intelligent decision-making. Thus, modern intelligent decision support systems must not only be highly accurate but also possess strong interpretability (Das, Kim, and Chernova 2023).

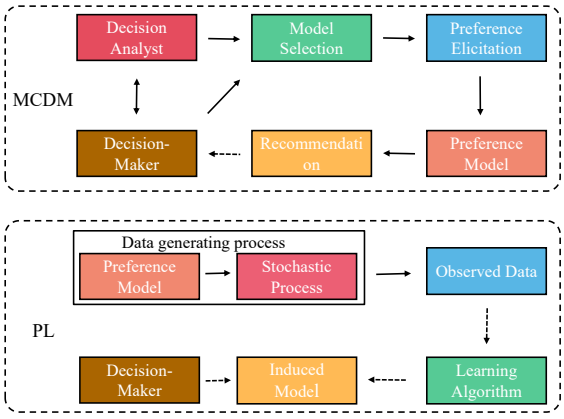


Figure 1: A simplified overview of MCDM and PL processes: MCDM involves active DM participation to construct preferences, whereas PL discovers preferences without further DM intervention.

Multi-criteria decision making (MCDM) and preference learning (PL) are two core fields in intelligent decision-making. MCDM aims to assist DMs in selecting the best alternative across multiple criteria (Zhao, Wu, and Dong 2024; Herin, Perny, and Sokolovska 2024). PL focuses on establishing rankings based on observations that reveal the preferences of individuals or groups. This ranking defines a binary preference relation that is irreflexive, antisymmetric, and transitive (Hüllermeier and Fürnkranz 2013; Fahandar and Hüllermeier 2018). Although MCDM and PL have evolved within different communities, they share the common objective of constructing practical decision models (Hüllermeier and Słowiński 2024a). Consequently, MCDM and PL can complement and enhance each other to some extent (Herin, Perny, and Sokolovska 2023; Liao, Liao, and Zhang 2023; Yijun, Mengzhuo, and Qingpeng 2023; Liu, Kadziński, and

*Corresponding author
Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Liao 2023). However, combining MCDM and PL to achieve the aforementioned goals remains a challenge, especially when dealing with large amounts of data and complex criteria that interdependently interact. This is primarily attributed to the inherent flaws of most existing methods, including the following key aspects:

i) **Capturing the relationships between decision factors.** In contemporary decision-making scenarios, the relationships among decision factors (including alternatives and criteria) are often complex. Single decision analysis tools usually fail to adequately reflect the interactions between these factors. Although some studies attempt to address this (Corrente et al. 2013; Wu et al. 2023), they often only handle problems with fewer decision factors due to limitations in the dimensions of relationship information.

ii) **Parsing the preferences of DMs.** The preferences of DMs themselves may not be entirely clear or fixed, but can be adjusted continually based on new information and experiences. Many models struggle to quantify such dynamic preferences due to limitations in the decision-making environment (Aggarwal and Fallah Tehrani 2019; Laidlaw and Russell 2021), resulting in final outcomes that do not align with the subjective perceptions of the DMs.

To tackle these constraints, we introduce the concepts of graph models and large language models (LLMs) into intelligent decision-making, utilizing graph structures to model the relationships between decision factors and employing LLMs to parse the preferences of DMs. i) Graphs are a very natural way to represent relationships, typically used to illustrate interactions between objects. Building on this foundation, graph neural networks (GNNs) have emerged as a powerful machine learning framework for processing graph-structured data (Scarselli et al. 2008). Their popularity stems from the ability to directly operate on the graph, learning node representations by leveraging the inherent connectivity of the graph (Kipf and Welling 2016; Veličković et al. 2017; Xu et al. 2018). If intelligent decision-making problems are mapped onto the graph domain, then GNNs can be used to update the features of these factors. By leveraging the expressive capability of GNNs, the complex relationships between decision factors are captured, thereby improving the predictive accuracy of the decision model. ii) LLMs, such as GPT-4 (Achiam et al. 2023) and LLaMA (Touvron et al. 2023), have recently gained widespread attention within the natural language processing (NLP) community and other fields, which opens up new possibilities for intelligent decision-making. If LLMs are integrated into intelligent decision systems, they can provide high-quality textual descriptions of DMs' preferences. This proficiency allows the model to comprehend and process complex language inputs effectively, enabling it to extract information quickly. By parsing and quantifying DMs' preferences, LLMs can boost the model's interpretability and transparency.

In light of the aforementioned insights, we propose an approach termed **I**ntelligent **D**ecision-Making with Integrated **G**raph **M**odels and Large **L**anguage **M**odels (**ID-GMLM**). The objective of this research is to refine and validate the ID-GMLM framework to ensure it not only captures, integrates, and models the complex relational dynamics among decision factors but also effectively translates these dynamics into actionable insights that align with DMs' preferences. Overall, the main contributions of this work are summarized as follows:

- We design a multi-task learning architecture in which different GNN structures are constructed for the primary and auxiliary tasks. This setup is intended to capture the interactions among alternatives and to model the mutual influences among criteria.
- We introduce LLMs to parse the criterion descriptions provided by DMs, and transform these descriptions into criterion weights. These weights serve as crucial inputs in the model's decision-making process.
- We conduct experiments on decision-making benchmarks and demonstrate the effectiveness and superiority of the proposed approach through empirical analysis.

Related Work

Combining MCDM and PL

MCDM and PL belong to two distinct research areas within intelligent decision-making, but both aim to construct practical decision models and infer model parameters from holistic judgments. It is precisely this commonality and difference that promote cross-research between them, leading to increasing attention on combining mathematical preference modeling with machine learning model identification techniques (Hüllermeier and Słowiński 2024b). Most related work starts with the preference model from MCDM and modifies it using machine learning algorithms. NEUR-HCI (Bresson et al. 2021), Sugeno Classifier (Abbaszadeh and Hüllermeier 2020), ELECTRE TRI-rC (Kadziński and Szczepański 2022), PLMCS (Liu et al. 2021), and ANN-UTADIS (Martyn and Kadziński 2023) all belong to such a direction. They use the MCDM model as the core framework, with PL methods as auxiliary tools to enhance performance. The advantage of this strategy is that it can estimate DMs' preferences from data while ensuring the rationality of decisions (Fu et al. 2021; Liao et al. 2023; Wu, Liao, and Zhang 2024). However, these approaches focus on static decision environments and fail to align with DMs' dynamic preferences. Furthermore, they often struggle to produce reasonable outcomes in decisions involving relational information between factors.

Uses of GNNs and LLMs in Decision-Making

GNNs offer a new perspective on decision-making by processing and analyzing graph data. Consequently, researchers have attempted to apply GNNs in intelligent decision-making, leveraging their expressive power to enhance decision model performance. For example, in the field of safety decision-making, (da Silva, Pedrini, and Santos 2023) introduced a universal network topology learning method for any GNN framework. (Zhang et al. 2023) emphasized the decision-making role of GNNs in social recommendation issues and designed a graph learning-augmented heterogeneous GNN. (Li et al. 2023) developed a hierarchical GNN

method for predicting patient treatment preferences in medical decision-making problems. It can be observed that GNNs have unique value in various decision-making domains, and their application is redefining how to understand and optimize the decision-making process (Dong et al. 2022). However, many methods do not fully meet the prerequisites of intelligent decision-making discussed in this paper. Most focus on the decision-making role of GNNs in different issues, rather than the inherent connection between GNNs and the final decision.

As a significant technology in NLP, LLMs have demonstrated great potential in assisting and enhancing decision-making (Chen et al. 2024). Integrating LLMs with DM preferences is a key direction for the application of LLMs in decision-making, and scholars have conducted many in-depth studies in this area (Handa et al. 2024; Liu et al. 2023; Korbak et al. 2023; Muldrew et al. 2024; Wang et al. 2024; Song et al. 2024). These models collectively aim to enhance the alignment of LLMs with human preferences in PL. The common theme across them is to improve preference elicitation and content generation to better reflect DMs' preferences. Although they focus on flexible decision scenarios, they demonstrate the rationale and feasibility of applying LLMs in the field of intelligent decision-making.

Methodology

Problem Statement

Consider a set of alternatives $A = \{a_1, a_2, \ldots, a_m\}$, each alternative a_i $(i = 1, \ldots, m)$ is assigned to a predefined set of decision classes $Cl = \{Cl_1, Cl_2, \ldots, Cl_q\}$, such that Cl_{s+1} is preferred over Cl_s (denoted by $Cl_{s+1} \succ Cl_s$), $s = 1, 2, \ldots, q-1$. $C = \{c_1, c_2, \ldots, c_n\}$ is a finite set of n evaluation criteria, where $c_j(a_i)$ $(j = 1, 2, \ldots, n)$ represents the performance of alternative a_i on criterion c_j, and all criteria are of the benefit type, meaning that for any $a_i \in A$, the higher the $c_j(a_i)$, the more preferred a_i performs under criterion c_j. In this paper, our task is to develop a decision model using the aforementioned information, which can correctly classify unknown alternatives and provide explanations for the results.

GNNs-Based Multi-Task Learning Strategy

To capture the complex relational information between decision factors, we first map the intelligent decision-making problem onto the graph domain and construct two specific graphs: the alternative preference graph (G_p) and the criterion relationship graph (G_c). In constructing these graphs, we leverage the k-nearest neighbors approach to establish connections between nodes based on their similarity. Specifically, for G_p, if the k-nearest neighbors of alternative a_i are denoted as $\mathcal{N}_i$, the similarity measure between alternatives is defined as follows:

$$\tilde{\mathbf{A}}_{ij} = \begin{cases} \exp(\|\mathbf{x}_i - \mathbf{x}_j\|^2/\sigma^2), & j \in \mathcal{N}_i \\ 1, & j = i \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

where $\tilde{\mathbf{A}} \in \mathbb{R}^{m \times m}$ is the similarity matrix with m alternatives, $\mathbf{x}_i$ and $\mathbf{x}_j$ are feature vectors of alternatives a_i and a_j, represented as $\mathbf{x}_i = \{c_1(a_i), c_2(a_i), \ldots, c_n(a_i)\}$, $\mathbf{x}_j = \{c_1(a_j), c_2(a_j), \ldots, c_n(a_j)\}$. Since the similarity matrix contains pairwise similarity values, we transpose $\tilde{\mathbf{A}}$ to obtain a symmetric similarity matrix $\mathbf{A}_p$, which serves as the adjacency matrix for G_p.

$$\mathbf{A}_p = \frac{\tilde{\mathbf{A}} + \tilde{\mathbf{A}}^T}{2}. \tag{2}$$

$G_p = (\mathbf{A}_p, \mathbf{X}_p)$, where $\mathbf{X}_p \in \mathbb{R}^{m \times n}$ is the alternative feature matrix. $G_c = (\mathbf{A}_c, \mathbf{X}_c)$ is a fully connected graph with n criteria, where $\mathbf{A}_c \in \mathbb{R}^{n \times n}$ is the adjacency matrix and $\mathbf{X}_c \in \mathbb{R}^{n \times m}$ is the criterion feature matrix.

Based on the two constructed graphs, we design an alternative preference network (APN) and a criterion relationship network (CRN) within the multi-task learning framework using GNNs. The former belongs to the primary task, which captures relationships between alternatives, while the latter is an auxiliary task aimed at modeling interactions among criteria. The APN is a GNN with residual connections, which aggregates and updates node features in G_p via graph convolutions.

$$\mathbf{X}_p^{(l+1)} = \mathrm{ReLU}(\mathbf{D}_p^{-\frac{1}{2}}\mathbf{A}_p\mathbf{D}_p^{-\frac{1}{2}}\mathbf{X}_p^{(l)}\mathbf{W}_p^{(l+1)} + \mathbf{X}_p^{(l)}\mathbf{W}_{sc}^{(l+1)}), \tag{3}$$

where $\mathbf{D}_p$ is the degree matrix of $\mathbf{A}_p$, $\mathbf{X}_p^l$ is the output of the l-th layer, and $\mathbf{W}_p^{l+1}$ is the weight matrix of the (l+1)-th layer. $\mathbf{X}_p^{(l)}\mathbf{W}_{sc}^{(l+1)}$ represents the residual connection, which includes two cases: if the input and output dimensions match, it is an identity function; otherwise, it performs a linear transformation. The CRN operates on G_c and consists of two layers of graph convolution. The input is the criterion feature matrix $\mathbf{X}_c$, and the output is the criterion relationship matrix $\mathbf{R}_c$.

$$\mathbf{X}_c^{(l+1)} = \mathrm{ReLU}(\mathbf{D}_c^{-\frac{1}{2}}\mathbf{A}_c\mathbf{D}_c^{-\frac{1}{2}}\mathbf{X}_c^{(l)}\mathbf{W}_c^{(l+1)}). \tag{4}$$

LLMs-Based Criterion Weights Generation

To integrate the intuition and experience of DMs into data-driven decision models, we introduce LLMs to parse and quantify the DMs' preferences. Since it is often easier for DMs to describe criteria qualitatively rather than assign explicit numerical weights, we use these natural language descriptions to derive criterion weights that reflect the relative importance of each criterion (Eigner and Händler 2024).

We first define a set of decision problem descriptions as $\mathcal{P}$, which includes the problem's background introduction and a part of the original decision data, serving as the prompt input for the LLM. Within $\mathcal{P}$, we frame the task as follows: "*You serve as an intelligent assistant that helps me assess the potential importance of each criterion in decision problem. I will provide you with information about a portion of the original decision examples and preferences for each criterion.*" Next, we define a set of criteria descriptions $\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_n\}$, where $\mathcal{D}_i$ is a natural language explanation of the i-th criterion provided by the DM. For example, the criteria might be described as follows: "*Here are our preferences for each criterion: ['Criterion 1': 'This criterion assesses..., which is an extremely important factor.'].*"

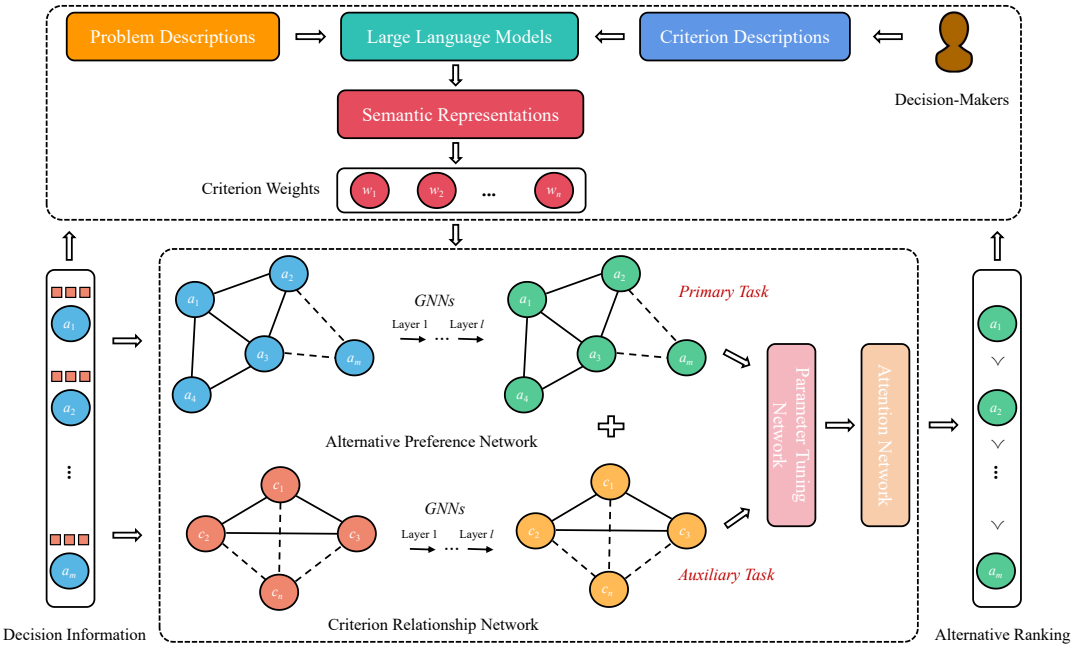


Figure 2: The ID-GMLM framework consists of the following components: The original decision information, including alternatives and criteria, is constructed into graphs and utilized in both the primary and auxiliary tasks involving GNNs. Based on problem descriptions and criterion descriptions provided by DMs, LLMs output the criterion weights, which are then applied in the decision-making process.

By providing the LLM with both $\mathcal{P}$ and $\mathcal{D}$ as input, we obtain a set of weight profiles for each criterion, which convey the perceived importance of the criterion. For instance, the LLM might produce a profile like: *"['Criterion 1': 'This criterion is critical due to...']*" These profiles are then encoded into semantic representations of criterion weights as:

$$\mathbf{W} = \tau(LLMs(\mathcal{P}, \mathcal{D})), \tag{5}$$

where $\tau(\cdot)$ is a text embedding model (Su et al. 2022; Ren et al. 2024) that converts the textual output into fixed-length vectors to retain their inherent meaning.

Parameter Tuning Network

To dynamically reflect the preferences of DMs, we design a parameter tuning network based on criterion weights. In this setup, $\mathbf{W}$ is used either for initialization or as input to the parameter tuning network, which further refines and adjusts these weights. The network adopts a two-layer neural network structure, with criterion weights $\mathbf{W} \in \mathbb{R}^{n\times 1}$ as the input and tuning factors $\mathbf{T} \in \mathbb{R}^{n\times 1}$ as the output.

$$\mathbf{T} = \text{Sigmoid}(\text{ReLU}(\mathbf{W}\mathbf{W}_{t_1} + \mathbf{b}_{t_1})\mathbf{W}_{t_2} + \mathbf{b}_{t_2}), \tag{6}$$

where $\mathbf{W}_{t_1}$ and $\mathbf{W}_{t_2}$ are the weight matrices of the two-layer network, and $\mathbf{b}_{t_1}$ and $\mathbf{b}_{t_2}$ are the biases, respectively. The tuned weights $\mathbf{T}$ are applied to the input feature $\mathbf{X}_p$ of the decision model, which adjusts the influence of different features in the decision-making process.

$$\mathbf{X}_{tp} = \mathbf{X}_p \odot \mathbf{T}. \tag{7}$$

Here, $\mathbf{X}_{tp}$ represents the tuned alternative feature matrix, which is then input into the APN to obtain the embedding representation of the alternative.

Attention Network

To better balance the primary and auxiliary tasks and ensure a reasonable weight distribution, we introduce an attention network. The structure of this network is similar to that of the parameter tuning network, as both are shallow neural networks. The input to the attention network is the embedding representation from the APN, and the output is a weight vector.

$$\boldsymbol{\alpha} = \sigma(\text{APN}(G_p, \mathbf{X}_{tp})\mathbf{W}_\alpha + \mathbf{b}_\alpha), \tag{8}$$

where σ represents the activation function, and $\boldsymbol{\alpha} \in \mathbb{R}^{m\times 1}$ denotes the attention values used to measure the proportion of loss weights for the auxiliary task in multi-task learning.

Optimization Objective

We divide the optimization objective of the decision model into three parts: the primary task, the auxiliary task, and the regularization term. Assuming there are M alternative pairs, for each pair, the loss $\mathcal{L}_p$ between the predicted output $\hat{y}_i$ and the true label y_i is defined as follows:

$$\mathcal{L}_p = -\frac{1}{M}\sum_{i=1}^{M}[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)]. \tag{9}$$

Here, $\hat{y}_i$ is obtained by subtracting the output embeddings of the APN and mapping it to the [0,1] range. y_i is derived by comparing the decision classes Cl of alternative pairs.

Let $\mathbf{R}_c^{'}$ be the target relationship matrix, and define the loss $\mathcal{L}_c$ as:

$$\mathcal{L}_c = \frac{1}{n^2}\sum_{p=1}^{n}\sum_{q=1}^{n}(\mathbf{R}_c[p,q] - \mathbf{R}_c^{'}[p,q])^2, \tag{10}$$

where n is the number of criteria. $\mathbf{R}_c^{'}$ is constructed by computing the correlation ρ_j between each criterion feature $c_j(a_i)$ of alternative a_i in $\mathbf{X}_p$ and the decision class Cl^{a_i}:

$$\rho_j = \frac{\mathrm{cov}(c_j(a_i), Cl^{a_i})}{\sigma_{c_j(a_i)}\sigma_{Cl^{a_i}}}, \tag{11}$$

where cov denotes covariance and σ represents standard deviation. The target matrix is then defined as $\mathbf{R}_c^{'}[p,q] = \rho_p \cdot \rho_q$.

To prevent overfitting and ensure consistency between model weights and DM weights, we employ L2 regularization. The loss $\mathcal{L}_r$ is defined as:

$$\mathcal{L}_r = \lambda \cdot \|\mathbf{W}_p - \mathbf{W}\|_2^2, \tag{12}$$

here λ is the regularization coefficient, and $\mathbf{W}_p$ is the original weight parameter in the APN, which is initialized as a vector of ones, equal in dimension to the number of criteria.

Combining the above three types of losses, the total loss is expressed as follows:

$$\mathcal{L} = \mathcal{L}_p + \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{\alpha}_i \mathcal{L}_{c,i} + \mathcal{L}_r, \tag{13}$$

where m is the number of alternatives, and $\boldsymbol{\alpha}_i$ is the attention weight for the i-th alternative (node).

Experiments

In this section, we conduct experiments to address the following research questions:

- **RQ1:** To what extent does ID-GMLM improve the decision effectiveness over other intelligent decision models?
- **RQ2:** How does each component in ID-GMLM contribute to enhancing decision-making accuracy?
- **RQ3:** Does incorporating LLMs into ID-GMLM improve the interpretability of decision-making?
- **RQ4:** How do key parameters influence the performance of ID-GMLM?

Experimental Setup

Datasets We collect six publicly available datasets from the UCI[1] and WEKA[2] repositories, which are summarized in Table 1. These datasets are selected because they meet two essential conditions: first, the outputs are measured on an ordered categorical scale; second, all criteria have a monotonic influence on the ranking of alternatives.

Baselines Since comparisons between MCDM and PL are primarily methodological and conceptual, rather than experimental (Hüllermeier and Słowiński 2024b), we select four types of MCDM and PL combination methods as baselines for ID-GMLM. Brief descriptions of these methods are provided below:

- **NEUR-HCI** (Bresson et al. 2021): A machine learning-based approach for the automatic recognition of hierarchical MCDM models.

[1]http://archive.ics.uci.edu/ml/.

[2]http://www.cs.waikato.ac.nz/ml/weka/datasets.html.

Dataset	Alternatives	Criteria	Classes	Comparisons
DBS	120	8	2	7140
BCC	286	7	2	40,755
MPG	392	7	3	76,636
ESL	488	4	4	118,828
LEV	1000	4	3	499,500
CEV	1728	6	4	1,492,128

Table 1: Statistics of the six datasets used in experiments.

- **ELECTRE TRI-rC** (Kadziński and Szczepański 2022): A crossover method between MCDM and PL that incorporates a single characteristic profile to describe each decision class.
- **PLMCS** (Liu et al. 2021): A hybrid approach that merges MCDM and PL, using an additive piecewise-linear value function as its basic preference model.
- **ANN-UTADIS** (Martyn and Kadziński 2023): A PL algorithm that utilizes neural networks to infer threshold-based sorting parameters from assignment examples.

Evaluation Metrics We adopt widely used protocols: Normalized Discounted Cumulative Gain (NDCG), Average Precision (AP), and C-index to evaluate the performance of the proposed approach. NDCG and AP emphasize accuracy at the top of rankings, awarding higher scores for correct top predictions (He et al. 2020). C-index measures the overall consistency between predicted and actual rankings, not just the top few (Aggarwal and Fallah Tehrani 2019).

Parameters Settings We split the data into training, validation, and test sets in a 7:2:1 ratio. Our model consists of a 2-layer residual GNN with a mean aggregator and hidden layers sized 2–5 times the input layer, alongside a 2-layer fully connected GCN with hidden layers three times the input size. We perform hyperparameter tuning with learning rates and regularization weights from {1e-1, 1e-2, 1e-3, 1e-4, 1e-5}, weight decay values {1e-3, 1e-4}, and k-nearest neighbors ranging from 2 to 10. Training uses the Adam optimizer for up to 500 epochs with early stopping, and initial DM preferences are based on dataset descriptions. We employ *gpt-3.5-turbo* and `text-embedding-ada-002` models to obtain semantic representations of criterion weights. Each model is run 100 times on an Intel(R) Xeon(R) Gold 5218R CPU @ 2.10GHz with 96GB memory, and we report the average results.

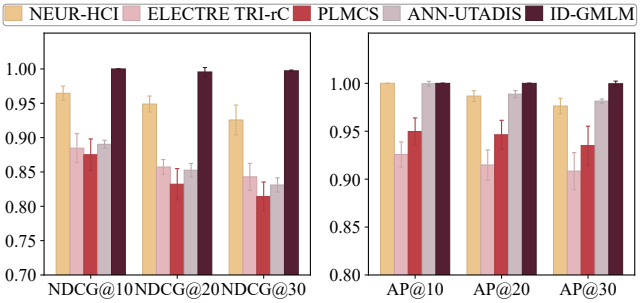
Overall Performance (RQ1)

We compare the proposed ID-GMLM with other popular approaches, and the results are summarized in Table 2. The following observations can be made: i) ID-GMLM outperforms other baselines on most datasets. Specifically, for NDCG, AP, and C-index, ID-GMLM achieves maximum relative improvements of 20.3%, 25.2%, and 24.2%, respectively on the BCC dataset. The results demonstrate the effectiveness of the proposed approach. ii) ID-GMLM shows more stable performance across different datasets compared to

Dataset	Metrics	NEUR-HCI	ELECTRE TRI-rC	PLMCS	ANN-UTADIS	ID-GMLM
DBS	NDCG	0.9540 ± 0.0028	0.9607 ± 0.0083	0.9199 ± 0.0018	0.9882 ± 0.0247	$\mathbf{0.9958 \pm 0.0016}$
	AP	0.9512 ± 0.0246	0.9422 ± 0.0160	0.8970 ± 0.0394	0.9617 ± 0.0249	$\mathbf{0.9672 \pm 0.0110}$
	C-index	0.9394 ± 0.0179 (4)	0.9464 ± 0.0245 (3)	0.9071 ± 0.0166 (5)	$\mathbf{0.9718 \pm 0.0121}$ (1)	0.9662 ± 0.0094 (2)
BCC	NDCG	0.9375 ± 0.0014	0.8476 ± 0.0013	0.8295 ± 0.0433	0.8311 ± 0.0026	$\mathbf{0.9979 \pm 0.0008}$
	AP	0.8944 ± 0.0361	0.7817 ± 0.0270	0.7941 ± 0.0505	0.7889 ± 0.0541	$\mathbf{0.9791 \pm 0.0080}$
	C-index	0.9256 ± 0.0373 (2)	0.8148 ± 0.0102 (3)	0.8065 ± 0.0349 (4)	0.8027 ± 0.0235 (5)	$\mathbf{0.9975 \pm 0.0011}$ (1)
MPG	NDCG	0.9879 ± 0.0006	0.9349 ± 0.0082	0.9311 ± 0.0152	0.9880 ± 0.0136	$\mathbf{0.9985 \pm 0.0004}$
	AP	0.9588 ± 0.0394	0.8872 ± 0.0267	0.9075 ± 0.0328	0.9649 ± 0.0413	$\mathbf{0.9838 \pm 0.0043}$
	C-index	0.9617 ± 0.0393 (2)	0.9067 ± 0.0425 (3)	0.8986 ± 0.0513 (5)	0.9522 ± 0.0275 (4)	$\mathbf{0.9902 \pm 0.0033}$ (1)
ESL	NDCG	0.9758 ± 0.0060	0.9856 ± 0.0147	0.9657 ± 0.0107	0.9986 ± 0.0022	$\mathbf{0.9996 \pm 0.0002}$
	AP	0.9603 ± 0.0361	0.9692 ± 0.0337	0.9482 ± 0.0319	0.9815 ± 0.0219	$\mathbf{0.9957 \pm 0.0029}$
	C-index	0.9672 ± 0.0330 (4)	0.9680 ± 0.0317 (3)	0.9463 ± 0.0446 (5)	0.9913 ± 0.0115 (2)	$\mathbf{0.9976 \pm 0.0008}$ (1)
LEV	NDCG	0.9856 ± 0.0043	0.9274 ± 0.0091	0.9401 ± 0.0097	0.9779 ± 0.0148	$\mathbf{0.9978 \pm 0.0025}$
	AP	0.9679 ± 0.0308	0.8988 ± 0.0462	0.9182 ± 0.0458	0.9683 ± 0.0430	$\mathbf{0.9759 \pm 0.0300}$
	C-index	0.9719 ± 0.0244 (2)	0.9065 ± 0.0294 (5)	0.9256 ± 0.0410 (4)	0.9425 ± 0.0364 (3)	$\mathbf{0.9839 \pm 0.0137}$ (1)
CEV	NDCG	0.9896 ± 0.0083	0.9802 ± 0.0145	0.9829 ± 0.0059	0.9814 ± 0.0106	$\mathbf{0.9962 \pm 0.0018}$
	AP	$\mathbf{0.9512 \pm 0.0206}$	0.9197 ± 0.0335	0.9266 ± 0.0346	0.9467 ± 0.0471	0.9449 ± 0.0220
	C-index	0.9835 ± 0.0330 (2)	0.9586 ± 0.0396 (4)	0.9520 ± 0.0183 (5)	0.9766 ± 0.0210 (3)	$\mathbf{0.9960 \pm 0.0026}$ (1)
Average Rank		2.67	3.50	4.67	3.00	1.17

Table 2: Performance of different approaches in terms of three evaluation metrics (mean $\pm$ standard deviation), with the best results in bold and the second-best results in underline. Rankings are in parentheses, where (1) represents the best and (5) the worst.

other baselines. For example, the average C-index for baselines on the ESL dataset is around 96.8%, while ID-GMLM reaches 99.8%; and for the BCC dataset, the average C-index for baselines drops to about 83.7%, while ID-GMLM still achieves 99.7%. These results highlight the superiority of the proposed approach. Figure 3 shows the changes in NDCG@n and AP@n values for various approaches on the BCC and LEV datasets. As can be seen, ID-GMLM consistently performs the best. On the BCC dataset, its NDCG@10 is nearly 14.3% higher than that of the poorest-performing approach. On the LEV dataset, its AP@30 is nearly 10.0% higher than than that of the poorest-performing approach.



(a) BCC (b) LEV

Figure 3: Ranking accuracy of all approaches in terms of NDCG@n and AP@n.

Ablation Study (RQ2)

To analyze the contribution of each component to decision accuracy, we conduct ablation studies on ID-GMLM by removing key components, creating three variants: ID-GMLM$_{\text{w/o CRN}}$ (no CRN), ID-GMLM$_{\text{w/o PTN}}$ (no parameter tuning network), and ID-GMLM$_{\text{w/o AN}}$ (no attention network). Figure 4 displays the results of the ablation analysis. The following conclusions can be drawn: i) The performance of ID-GMLM consistently outperforms the other three variants, indicating the effectiveness of using CRN, parameter tuning network, and attention network simultaneously. ii) The performance of ID-GMLM$_{\text{w/o CRN}}$ is the poorest, highlighting CRN's critical role in analyzing relationships between criteria. iii) The performances of ID-GMLM$_{\text{w/o PTN}}$ and ID-GMLM$_{\text{w/o AN}}$ are relatively close and both slightly inferior to ID-GMLM. We believe this is because ID-GMLM$_{\text{w/o PTN}}$ lacks the capability to directly tune feature weights in the APN. ID-GMLM$_{\text{w/o AN}}$ does not adjust the importance of the primary and auxiliary tasks, but since the model's output mainly relies on the former, its impact on the final performance is not particularly significant.

Efficiency Analysis (RQ3)

To explore the effectiveness of LLMs, we analyze the impact of different types of criterion weight inputs on decision outcomes. The experimental results are shown in Table 3: i) LLM_O (original weights by LLMs) outperforms LLM_M (equal weights by LLMs) and LLM_R (random weights by LLMs). Specifically, LLM_O improves NDCG, AP, and C-index by 0.27%, 0.55%, and 0.56% over LLM_M, and by 0.34%, 0.81%, and 0.69% over LLM_R, demonstrating LLM's reliability in ID-GMLM. ii) Although using LLM_M

	LLM_O			LLM_M			LLM_R		
Metrics	DBS	BCC	MPG	DBS	BCC	MPG	DBS	BCC	MPG
NDCG	0.9971	0.9988	0.9989	0.9926	0.9968	0.9967	0.9911	0.9975	0.9980
AP	0.9767	0.9874	0.9884	0.9689	0.9780	0.9860	0.9647	0.9746	0.9789
C-index	0.9762	0.9985	0.9930	0.9635	0.9959	0.9875	0.9644	0.9969	0.9862
Metrics	ESL	LEV	CEV	ESL	LEV	CEV	ESL	LEV	CEV
NDCG	0.9993	0.9993	0.9977	0.9965	0.9986	0.9944	0.9957	0.9931	0.9949
AP	0.9947	0.9938	0.9683	0.9942	0.9837	0.9647	0.9879	0.9823	0.9630
C-index	0.9955	0.9935	0.9977	0.9882	0.9876	0.9937	0.9871	0.9838	0.9927

Table 3: Performance comparison of ID-GMLM with different criterion weight inputs generated by LLMs.

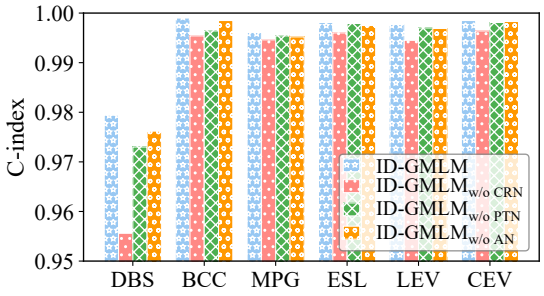
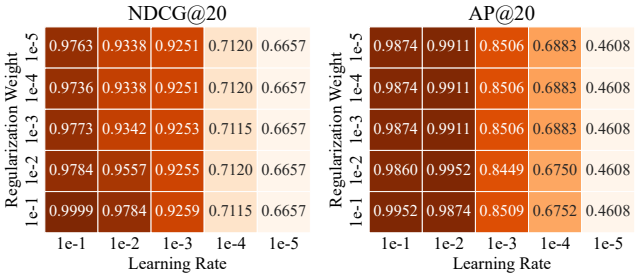


Figure 4: Comparison of the performance of different variants of ID-GMLM.



(a) DBS (b) BCC

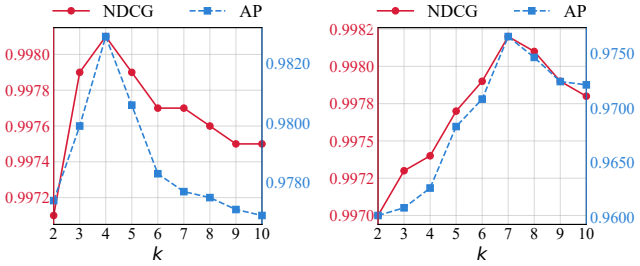
Figure 5: The performance of ID-GMLM with different learning rates and regularization weights.

and LLM_R leads to some performance decline, the overall magnitude is minimal. This is mainly due to the robustness of ID-GMLM, which maintains relatively stable performance under different criterion weight configurations. However, this does not diminish the contribution of the weights generated by LLMs to the model.

Sensitivity Analysis (RQ4)

Regularization Weights & Learning Rates Figure 5 records the performance of ID-GMLM under different combinations of learning rates and regularization weights on DBS and BCC datasets in terms of NDCG@20 and AP@20. Key observations include: i) The model performs stably at learning rates of {1e-1, 1e-2} and performs well across all regularization weights ii) Performance remains largely unchanged with different regularization weights at the same learning rate but decreases as the learning rate is reduced, indicating the model is less sensitive to regularization weights and more sensitive to learning rates.

***k*-Nearest Neighbors** Figure 6 examines the effect of varying the number of neighbors k (from 2 to 10) on ID-GMLM's performance measured by NDCG and AP. We find that: i) Optimal performance occurs when k is between 3 and 7. ii) As k increases, both NDCG and AP values initially rise and then fall. This may happen because with a small k, only limited information is propagated from the nearest neighbors, leading to poorer performance. Conversely, a very large k can smooth out features or introduce noise, causing performance to plateau and then decline.



(a) MPG (b) CEV

Figure 6: Sensitivity analysis of the parameter k for ID-GMLM.

Conclusion

In this paper, we explore the effective combination of MCDM and PL, proposing an Intelligent Decision-Making with Integrated Graph Models and Large Language Models (**ID-GMLM**). Specifically, we employ a multi-task learning framework to derive the ranking of alternatives, while incorporating GNNs and LLMs to capture relationships between decision factors and learn DM preferences, respectively. Experimental results on six benchmarks demonstrate that our approach is outperforms existing models. Our future work will focus on better aligning DM preferences with LLM outputs, thereby bridging the gap between DM knowledge and data-driven models.

Acknowledgements

This study was supported by the National Key Research and Development Program of China under Grant 2021YFB3300700.

References

Abbaszadeh, S.; and Hüllermeier, E. 2020. Machine learning with the sugeno integral: The case of binary classification. *IEEE Transactions on Fuzzy Systems*, 29(12): 3723–3733.

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Aggarwal, M.; and Fallah Tehrani, A. 2019. Modelling human decision behaviour with preference learning. *INFORMS Journal on Computing*, 31(2): 318–334.

Bresson, R.; Cohen, J.; Hüllermeier, E.; Labreuche, C.; and Sebag, M. 2021. Neural representation and learning of hierarchical 2-additive Choquet integrals. In *Proceedings of the 29th International Conference on International Joint Conferences on Artificial Intelligence*, 1984–1991.

Chen, J.; Lin, H.; Han, X.; and Sun, L. 2024. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 17754–17762.

Corrente, S.; Greco, S.; Kadziński, M.; and Słowiński, R. 2013. Robust ordinal regression in preference learning and ranking. *Machine Learning*, 93: 381–422.

da Silva, E. S.; Pedrini, H.; and Santos, A. 2023. Applying Graph Neural Networks to Support Decision Making on Collective Intelligent Transportation Systems. *IEEE Transactions on Network and Service Management*, 20(4): 4085–4096.

Das, D.; Kim, B.; and Chernova, S. 2023. Subgoal-based explanations for unreliable intelligent decision support systems. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, 240–250.

Dong, Y.; Liu, N.; Jalaian, B.; and Li, J. 2022. Edits: Modeling and mitigating data bias for graph neural networks. In *Proceedings of the ACM Web Conference 2022*, 1259–1269.

Eigner, E.; and Händler, T. 2024. Determinants of llm-assisted decision-making. *arXiv preprint arXiv:2402.17385*.

Fahandar, M. A.; and Hüllermeier, E. 2018. Learning to rank based on analogical reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Fu, C.; Xue, M.; Liu, W.; Xu, D.; and Yang, J. 2021. Data-driven preference learning in multiple criteria decision making in the evidential reasoning context. *Applied Soft Computing*, 102: 107109.

Fu, H.; Manogaran, G.; Wu, K.; Cao, M.; Jiang, S.; and Yang, A. 2020. Intelligent decision-making of online shopping behavior based on internet of things. *International Journal of Information Management*, 50: 515–525.

Handa, K.; Gal, Y.; Pavlick, E.; Goodman, N.; Andreas, J.; Tamkin, A.; and Li, B. Z. 2024. Bayesian preference elicitation with language models. *arXiv preprint arXiv:2403.05534*.

He, X.; Deng, K.; Wang, X.; Li, Y.; Zhang, Y.; and Wang, M. 2020. Lightgcn: Simplifying and powering graph convolution network for recommendation. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, 639–648.

Herin, M.; Perny, P.; and Sokolovska, N. 2023. Learning preference models with sparse interactions of criteria. In *Proceedings of the 32nd International Conference on International Joint Conferences on Artificial Intelligence*, 3786–3794.

Herin, M.; Perny, P.; and Sokolovska, N. 2024. Learning GAI-Decomposable Utility Models for Multiattribute Decision Making. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 20412–20419.

Hüllermeier, E.; and Fürnkranz, J. 2013. Preference learning and ranking. *Machine Learning*, 93: 185–189.

Hüllermeier, E.; and Słowiński, R. 2024a. Preference learning and multiple criteria decision aiding: differences, commonalities, and synergies–part I. *4OR*, 1–31.

Hüllermeier, E.; and Słowiński, R. 2024b. Preference learning and multiple criteria decision aiding: differences, commonalities, and synergies—part II. *4OR*, 1–37.

Kadziński, M.; and Szczepański, A. 2022. Learning the parameters of an outranking-based sorting model with characteristic class profiles from large sets of assignment examples. *Applied Soft Computing*, 116: 108312.

Kipf, T. N.; and Welling, M. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Korbak, T.; Shi, K.; Chen, A.; Bhalerao, R. V.; Buckley, C.; Phang, J.; Bowman, S. R.; and Perez, E. 2023. Pretraining language models with human preferences. In *International Conference on Machine Learning*, 17506–17533.

Laidlaw, C.; and Russell, S. 2021. Uncertain decisions facilitate better preference learning. *Advances in Neural Information Processing Systems*, 34: 15070–15083.

Li, Q.; Chen, L.; Cai, Y.; and Wu, D. 2023. Hierarchical graph neural network for patient treatment preference prediction with external knowledge. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 204–215.

Liao, H.; He, Y.; Wu, X.; Wu, Z.; and Bausys, R. 2023. Reimagining multi-criterion decision making by data-driven methods based on machine learning: A literature review. *Information Fusion*, 101970.

Liao, Z.; Liao, H.; and Zhang, X. 2023. A contextual Choquet integral-based preference learning model considering both criteria interactions and the compromise effects of decision-makers. *Expert Systems with Applications*, 213: 118977.

Liu, J.; Kadziński, M.; and Liao, X. 2023. Modeling contingent decision behavior: A Bayesian nonparametric preference-learning approach. *INFORMS Journal on Computing*, 35(4): 764–785.

Liu, J.; Kadziński, M.; Liao, X.; and Mao, X. 2021. Data-driven preference learning methods for value-driven multiple criteria sorting with interacting criteria. *INFORMS Journal on Computing*, 33(2): 586–606.

Liu, W.; Wang, X.; Wu, M.; Li, T.; Lv, C.; Ling, Z.; Zhu, J.; Zhang, C.; Zheng, X.; and Huang, X. 2023. Aligning large language models with human preferences through representation engineering. *arXiv preprint arXiv:2312.15997*.

Martyn, K.; and Kadziński, M. 2023. Deep preference learning for multiple criteria decision analysis. *European Journal of Operational Research*, 305(2): 781–805.

Muldrew, W.; Hayes, P.; Zhang, M.; and Barber, D. 2024. Active Preference Learning for Large Language Models. *arXiv preprint arXiv:2402.08114*.

Ren, X.; Wei, W.; Xia, L.; Su, L.; Cheng, S.; Wang, J.; Yin, D.; and Huang, C. 2024. Representation learning with large language models for recommendation. In *Proceedings of the ACM on Web Conference 2024*, 3464–3475.

Scarselli, F.; Gori, M.; Tsoi, A. C.; Hagenbuchner, M.; and Monfardini, G. 2008. The graph neural network model. *IEEE transactions on neural networks*, 20(1): 61–80.

Simeone, A.; Zeng, Y.; and Caggiano, A. 2021. Intelligent decision-making support system for manufacturing solution recommendation in a cloud framework. *The International Journal of Advanced Manufacturing Technology*, 112(3): 1035–1050.

Song, F.; Yu, B.; Li, M.; Yu, H.; Huang, F.; Li, Y.; and Wang, H. 2024. Preference ranking optimization for human alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 18990–18998.

Su, H.; Shi, W.; Kasai, J.; Wang, Y.; Hu, Y.; Ostendorf, M.; Yih, W.-t.; Smith, N. A.; Zettlemoyer, L.; and Yu, T. 2022. One embedder, any task: Instruction-finetuned text embeddings. *arXiv preprint arXiv:2212.09741*.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; and Bengio, Y. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903*.

Wang, J.; Wang, H.; Sun, S.; and Li, W. 2024. Aligning language models with human preferences via a bayesian approach. *Advances in Neural Information Processing Systems*, 36.

Wu, X.; Liao, H.; and Zhang, C. 2024. Preference disaggregation analysis for sorting problems in the context of group decision-making with uncertain and inconsistent preferences. *Information Fusion*, 101: 102014.

Wu, Z.; Song, Y.; Ji, Y.; Qu, S.; and Gong, Z. 2023. Data-driven distributionally robust support vector machine method for multiple criteria sorting problem with uncertainty. *Applied Soft Computing*, 149: 110957.

Xu, K.; Hu, W.; Leskovec, J.; and Jegelka, S. 2018. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*.

Yijun, L.; Mengzhuo, G.; and Qingpeng, Z. 2023. Data-driven Preference Learning Methods for Multiple Criteria Sorting with Temporal Criteria. *arXiv preprint arXiv:2309.12620*.

Zhang, Y.; Wu, L.; Shen, Q.; Pang, Y.; Wei, Z.; Xu, F.; Chang, E.; and Long, B. 2023. Graph learning augmented heterogeneous graph neural network for social recommendation. *ACM Transactions on Recommender Systems*, 1(4): 1–22.

Zhao, S.; Wu, S.; and Dong, Y. 2024. Managing non-cooperative behaviors and ordinal consensus through a self-organized mechanism in multi-attribute group decision making. *Expert Systems with Applications*, 240: 122571.