Information Sciences 754 (2026) 123712

Contents lists available at ScienceDirect

Information Sciences

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

Granular-ball-driven knowledge acquisition and information fusion via PROMETHEE in multi-source information systems

Lingwei Wei [a], Weirui Ye [a], Weihua Xu [a,*], Shuyin Xia [b]

[a] College of Artificial Intelligence, Southwest University, Chongqing, 400715, PR China
[b] College of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing, 400065, PR China

ARTICLE INFO

Keywords:
Multi-source
Granular-ball computing
Multi-source granular-ball
Granular computing
Feature selection
Information fusion

ABSTRACT

Feature selection and information fusion are critical tasks in multi-source information systems. Traditional granular-ball computing models often struggle to effectively process the inherently complex correlations and uncertainties within multi-source data. To address this challenge, this paper proposes a novel framework that integrates multi-source granular-ball computing, variable precision rough sets (VPRS), and the PROMETHEE decision-making method. Specifically, a neighborhood search algorithm is first designed to generate multi-source granular-balls (MSGBs). On this basis, a Multi-Source Granular-Ball Variable Precision Rough Set (MSGB-VPRS) model is constructed. To quantify multi-granular uncertainty, a Zentropy-based measure is employed to evaluate feature importance. Subsequently, the PROMETHEE II method is utilized to rank features according to their net preference flows, resulting in a multi-source granular-ball feature selection method (MSGBP). This framework is further extended into a multi-source granular-ball information fusion strategy (MSGB-IFS). Extensive experiments on 12 UCI datasets demonstrate that the proposed MSGBP method achieves optimal or highly competitive classification accuracy across various classifiers. Furthermore, the MSGB-IFS strategy significantly outperforms traditional fusion methods. These comprehensive results substantiate the effectiveness and superiority of the proposed framework in addressing feature selection and information fusion challenges within multi-source environments.

1. Introduction

In the era of big data, the proliferation of multi-source, high-dimensional information systems has introduced substantial data management challenges [1]. Feature selection, also known as attribute reduction, serves as a pivotal preprocessing step across diverse analytical domains, spanning data mining tasks such as cost-sensitive learning [2] and class-imbalance mitigation [3], as well as machine learning paradigms including clinical diagnostics [4] and text classification [5]. Ultimately, this process plays a vital role in enhancing model performance [6], reducing computational overhead [7], and improving model interpretability [8]. Since Pawlak introduced rough set theory [9], it has garnered significant attention for its ability to elucidate the intrinsic mechanisms of feature selection. In response to increasingly complex data forms and diverse data types, scholars have proposed various novel feature selection algorithms based on rough sets, expanding into fuzzy neighborhood spaces [10] and hierarchical structures [11].

In recent years, granular computing theory, particularly the development of granular-ball computing, has simulated the human cognitive process from coarse-grained to fine-grained reasoning by replacing the processing unit from “points” to “balls”, significantly enhancing the efficiency and robustness of rough set-based feature selection algorithms [12]. Traditionally, granular-ball generation

* Corresponding author.
Email addresses: weilingwei_swu@163.com (L. Wei), reanye233@gmail.com (W. Ye), chxuwh@gmail.com (W. Xu), xiasy@cqupt.edu.cn (S. Xia).

https://doi.org/10.1016/j.ins.2026.123712
Received 19 January 2026; Received in revised form 30 May 2026; Accepted 30 May 2026
Available online 1 June 2026

involves treating the entire dataset as a single initial hypersphere, followed by a recursive splitting process. Granular-balls that fail to meet a predefined purity threshold are iteratively divided until all sub-balls satisfy the homogeneity requirement, yielding a hierarchical, multi-granularity structure. The granular-ball based rough set model [13] has been proven to significantly improve the efficiency and noise resistance of feature selection. Its success in single-source information systems offers new ideas for handling large-scale, high-dimensional data. The versatility of this approach is further evidenced by its successful extension into other domains, such as mean-shift outlier detection [14].

While granular-ball computing has achieved notable success in single-source systems, real-world data inherently originates from multiple homogeneous or heterogeneous sources. When data from different sources share the same or similar attribute sets structurally, meaning the data table fields or feature dimensions across different sources remain consistent, we refer to such systems as homogeneous multi-source information systems. While this type of system ensures semantic consistency of attributes, there may still be issues such as differences in data distribution, sampling noise, or uneven data quality among different data sources. Current research in the field of multi-source information systems primarily focuses on information fusion techniques [15], while feature selection methods specifically designed for multi-source environments remain relatively scarce. Most existing methods convert multi-source systems into single-source systems through information fusion for processing, which inevitably leads to information loss. Concurrently, ongoing research advances continue to address the complexities of multi-source data fusion and dynamic feature selection in such environments, exploring graph-driven feature selection [16] and supervised incremental approaches [17]. Crucially, existing granular-ball generation methods are tailored specifically for single-source systems [18], making them fundamentally ill-equipped to construct granular balls or handle data representation in multi-source spaces. These traditional approaches recursively split the entire dataset from a single initial granular ball, operating under the assumption of a single, homogeneous data distribution. However, this global assumption inherently breaks down when applied to data originating from multiple distinct sources. To address this limitation, the proposed multi-source granular-ball establishes a unified representation structure. It simultaneously captures data across sources while maintaining cross-source consistency. By utilizing a data-adaptive neighborhood search strategy, the method effectively quantifies inter-ball overlap to manage inter-source discrepancies and noise.

Concurrently, multi-source information systems in the real world often involve ordered data, meaning there are preference or dominance relationships between attribute values. In multi-criteria decision analysis (MCDM), the Preference Ranking Organization Method for Enrichment Evaluations (PROMETHEE) [19] provides a robust outranking framework for resolving complex evaluation problems involving multiple, potentially conflicting criteria. Its core lies in defining preference functions to quantify the degree of superiority of one alternative (or feature) over another on each criterion. These preferences are then aggregated across all criteria to calculate the net preference flow for each alternative. The net preference flow comprehensively reflects the overall advantage of an alternative in all pairwise comparisons, thereby enabling a complete ranking of alternatives. PROMETHEE is renowned for its conceptual clarity, ease of understanding, and ability to handle both qualitative and quantitative criteria. It has been widely deployed in fields ranging from environmental assessment [20] to supply chain management [21]. More recently, its integration into feature selection tasks has shown tremendous potential, providing strong theoretical support for evaluating feature importance [22] and ranking attributes in complex ordered multi-source environments [23]. Therefore, integrating PROMETHEE with multi-source granular-ball computing offers a promising new pathway to overcome the limitations of existing single-source granular-ball methods and achieve effective feature selection in multi-source contexts.

To address the aforementioned research gaps, this paper proposes an innovative framework for multi-source ordered decision information systems, which integrates multi-source granular-balls, variable precision rough sets, and the PROMETHEE multi-criteria decision-making method. The main contributions of this paper are as follows:

1. We design a neighborhood search-based granular-ball generation method capable of constructing a unified data representation structure, namely multi-source granular-balls, for multi-source information systems. This method can adapt to data of arbitrary distributions and effectively quantify the degree of overlap between granular-balls, laying the foundation for subsequent processing.
2. We extend the traditional variable precision rough set model to the multi-source context, utilizing multi-source granular-balls for joint lower and upper approximations, thereby handling noise and uncertainty in multi-source data more flexibly.
3. We introduce Zentropy to describe the uncertainty in different information sources, providing a more comprehensive perspective for feature evaluation. Furthermore, the feature selection procedure incorporates the PROMETHEE II methodology. By constructing attribute evaluation matrices and preference matrices, we calculate the net preference flow of features, enabling preference-based feature ranking and selection. Simultaneously, we propose an information fusion strategy based on similar principles.
4. Through extensive experiments on 12 UCI benchmark datasets, we compare the performance of the proposed feature selection method and information fusion strategy against various other methods. The results demonstrate that our method achieves the best or near-best classification accuracy on most datasets, fully validating the effectiveness, generality, and stability of the proposed framework in multi-source environments.

The remainder of this paper is organized as follows. Section 2 introduces the fundamentals of granular-ball computing, variable precision rough sets, and the PROMETHEE method. Section 3 elaborates on a novel neighborhood search-based technique for generating multi-source granular-balls. Section 4 puts forward a generalized multi-source granular-ball variable precision rough set model, introduces Zentropy for uncertainty measurement, and on this basis, constructs a feature selection algorithm by integrating the PROMETHEE II method. Section 5 further extends this framework by proposing a multi-source granular-ball-based information fusion

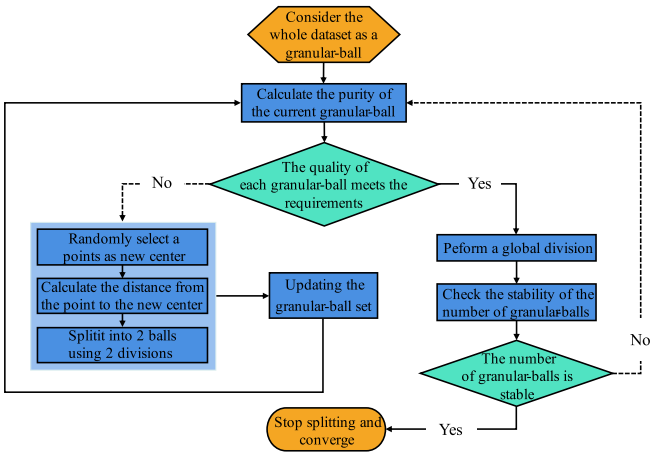


Fig. 1. Flowchart of the traditional generation method for granular-balls.

strategy for effective integration of multi-source data. Section 6 presents and analyzes the experimental results. Finally, Section 7 concludes the paper and outlines future research directions.

2. Preliminaries

This section introduces the fundamental concepts of granular-ball computing, variable precision rough sets and the PROMETHEE method.

2.1. Granular-ball rough set

The majority of established feature selection methods process information at the finest granularity, such as individual data points or pixels. However, this approach contrasts with human cognitive processes that typically evolve from coarse-grained to fine-grained reasoning. To address this gap, Xia et al. introduced Granular-Ball Computing (GBC), which replaces traditional point-based inputs with "granular-balls" [24]. Using granular-balls as fundamental units has two primary advantages. First, it enhances computational efficiency, particularly for large datasets. Second, it improves the effectiveness of rough set-based feature selection. This is because it facilitates the adaptive identification of neighborhood relationships, which leads to more robust approximations. In summary, granular-balls enhance both the efficiency and effectiveness of rough set-based feature selection algorithms.

Formally, given a dataset $U = \{x_1, x_2, \ldots, x_n\}$ with corresponding labels d_i for each sample x_i, a granular-ball gb is defined by a quintuple $(\mathring{U}, c, r, \mathring{d}, p)$. Here, $\mathring{U} \subseteq U$ denotes the subset of samples enclosed by the ball. The center c and radius r are computed as the sample mean and average distance to the center, respectively:

$$c = \frac{1}{|\mathring{U}|}\sum_{x_i \in \mathring{U}} x_i, r = \frac{1}{|\mathring{U}|}\sum_{x_i \in \mathring{U}} \|x_i - c\|. \tag{1}$$

The label $\mathring{d}$ of a granular-ball is defined as the majority vote of the labels belonging to the samples it contains. This is expressed as:

$$\mathring{d} = \mathrm{Mo}(\{d_i \mid x_i \in \mathring{U}\}). \tag{2}$$

The purity p, which quantifies the homogeneity of the ball, is defined as the proportion of samples possessing the majority label:

$$p = \frac{|\{x_i \mid d_i = \mathring{d}\}|}{|\mathring{U}|}. \tag{3}$$

Finally, the size of a granular-ball is simply the cardinality of its sample set, denoted by $|gb| = |\mathring{U}|$.

In the original algorithm [24], the generation process begins by treating the whole dataset as one initial granular-ball. This ball is then successively broken down through recursive partitioning into finer granular-balls. The iteration continues until the purity criterion is satisfied for every resulting ball, at which point the splitting process terminates. Fig. 1 displays a flowchart that outlines this primary method for generating granular-balls.

2.2. The variable precision rough set model

The classical rough set theory, pioneered by Pawlak, provides a fundamental framework for managing uncertainty and vagueness. However, its requirement for absolute set inclusion renders it sensitive to noise and inconsistencies commonly present in real-world data. To address this limitation, Ziarko introduced the Variable Precision Rough Set (VPRS) model, which incorporates a controlled level of misclassification tolerance.

Let U be a finite universe of discourse and A be a set of conditional attributes, which induce an equivalence relation R on U. For a target concept $X \subseteq U$, the classical lower and upper approximations of X are defined respectively as:

$$\underline{R}(X) = \{x \in U \mid [x]_R \subseteq X\}, \tag{4}$$

$$\overline{R}(X) = \{x \in U \mid [x]_R \cap X \neq \emptyset\}. \tag{5}$$

The VPRS model generalizes these notions by introducing a precision parameter $\beta \in [0, 0.5)$. This parameter represents the admissible classification error, allowing for a probabilistic interpretation of set inclusion. The conditional probability of an object x belonging to X, given its equivalence class $[x]_R$, is defined as:

$$P(X \mid [x]_R) = \frac{|[x]_R \cap X|}{|[x]_R|}. \tag{6}$$

Utilizing β, the β-lower approximation of X comprises objects whose equivalence classes are contained in X with a probability of at least $1 - \beta$:

$$\underline{R}_\beta(X) = \{x \in U \mid P(X \mid [x]_R) \geq 1 - \beta\}. \tag{7}$$

Conversely, the β-upper approximation includes objects for which the probability of belonging to X is strictly greater than β:

$$\overline{R}_\beta(X) = \{x \in U \mid P(X \mid [x]_R) > \beta\}. \tag{8}$$

The primary advantage of the VPRS model lies in its enhanced robustness and flexibility for analyzing imperfect data. It is noteworthy that the classical rough set model is a special case of the VPRS model when $\beta = 0$. Owing to these properties, VPRS has found successful applications in various fields such as data mining, feature selection, and decision support systems, where tolerating a certain degree of uncertainty is imperative.

2.3. The PROMETHEE algorithm

The Preference Ranking Organization METHod for Enrichment Evaluations (PROMETHEE) is a widely used outranking method in multi-criteria decision-making (MCDM) [25]. It evaluates and prioritizes alternatives through systematic pairwise comparisons. The method operates on a set of alternatives $\mathcal{A} = \{a_1, a_2, \ldots, a_n\}$ evaluated against a set of criteria $C = \{c_1, c_2, \ldots, c_m\}$, where each alternative a_i is characterized by its performance $f_k(a_i)$ on criterion c_k.

The PROMETHEE procedure begins by defining a preference function for each criterion. For any pair of alternatives (a_i, a_j), the preference strength of a_i over a_j with respect to criterion c_k is quantified by a preference function $P_k(a_i, a_j) = H_k(d_k(a_i, a_j))$, where $d_k(a_i, a_j) = f_k(a_i) - f_k(a_j)$ represents the performance difference, and $H_k(\cdot)$ is a non-decreasing function mapping this difference to a normalized preference degree between 0 and 1. Common forms of H_k include the usual, U-shape, V-shape, and Gaussian functions, accommodating different decision-maker preference structures.

These criterion-specific preferences are then aggregated into a global preference index $\pi(a_i, a_j) = \sum_{k=1}^{m} w_k \cdot P_k(a_i, a_j)$, which synthesizes the overall preference of a_i over a_j across all criteria. Here, w_k denotes the relative weight assigned to criterion c_k, subject to the normalization constraint $\sum_{k=1}^{m} w_k = 1$.

To establish a complete linear preorder of alternatives, the PROMETHEE II methodology computes three preference flows for each candidate. The positive outflow $\phi^+(a_i) = \frac{1}{n-1} \sum_{j \neq i} \pi(a_i, a_j)$ measures the average preference of alternative a_i over all others, indicating its overall strength. Conversely, the negative inflow $\phi^-(a_i) = \frac{1}{n-1} \sum_{j \neq i} \pi(a_j, a_i)$ captures the average preference of all other alternatives over a_i, reflecting its relative weakness. The net flow $\phi(a_i) = \phi^+(a_i) - \phi^-(a_i)$ then serves as a comprehensive performance metric, balancing both the strengths and weaknesses of each alternative.

The final ranking is obtained by ordering the alternatives in descending order of their net flows: if $\phi(a_{i_1}) > \phi(a_{i_2}) > \cdots > \phi(a_{i_n})$, then the complete preorder is given by $a_{i_1} \succ a_{i_2} \succ \cdots \succ a_{i_n}$. Renowned for its intuitive logic and flexibility in handling both quantitative and qualitative data, PROMETHEE has been successfully applied in diverse fields such as environmental management, supply chain optimization, and recently, as a ranking mechanism in feature selection algorithms.

3. Multi-source granular-balls

Traditional granular-ball computing is typically designed for single-source information systems, where all data originates from a uniform source, and is thus inadequate for data originating from multiple sources. To address this issue, this paper introduces an algorithm for generating granular-balls in multi-source information systems and proposes a definition of multi-source granular-balls. Unlike traditional granular-balls, the proposed multi-source granular-balls can represent data from multiple sources while ensuring consistency. This property makes them directly applicable for subsequent feature extraction in multi-source information systems. To illustrate the core concept intuitively, consider a medical diagnosis scenario where data is collected from multiple hospitals. Conventional methods would generate granular-balls separately for each hospital's patient data, which lacks cross-source correspondence. In contrast, the proposed multi-source granular-ball functions as a unified "medical record portfolio". It bundles one granular-ball from each hospital into a joint representational unit across institutions. This portfolio establishes a consistent structure among multi-source data, enabling unified and robust knowledge acquisition and information fusion. Its formal definition is provided below.

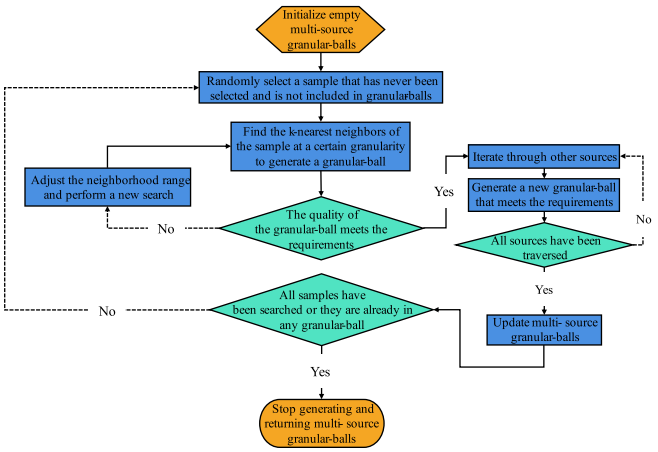


Fig. 2. Flowchart of the proposed neighborhood search-based method for generating multi-source granular-balls.

Definition 1. Let $\{DIS_j = (U, AT \cup D, V, f^j) | j = 1, 2, 3, \ldots, q\}$ be a multi-source decision information system, where $U = \{x_1, x_2, \ldots, x_m\}$ is the finite universe of samples, $AT = \{a_1, a_2, \ldots, a_n\}$ is the set of conditional attributes, D is the decision attribute, V is the domain of attribute values and $f^j : U \times (AT \cup D) \to V$ is the information function for the j-th source ($1 \le j \le q$). The multi-source granular-balls set $MSGB$ is defined as a collection of multi-source granular-balls $msgb_i$:

$$MSGB = \{msgb_1, msgb_2, \ldots, msgb_p\}, \tag{9}$$

$$msgb_i = \{gb_i^1, gb_i^2, \ldots, gb_i^q\}, \quad i \le p. \tag{10}$$

Here, $gb_i^j = (\mathring{U}_i^j, c_i^j, r_i^j, \mathring{d}_i^j, p_i^j) \in msgb_i$ denotes a single-source granular-ball, whose definition is described in Section 2.1, with j representing the source index. In contrast to the radius definition for standard granular-balls provided in Section 2.1, the radius r_i^j of a single granular-ball in a multi-source environment is defined as the average distance of two specific neighbor samples to the center:

$$r_i^j = \frac{\|x_a^j - c_i^j\| + \|x_b^j - c_i^j\|}{2}, \tag{11}$$

where x_a^j and x_b^j are the $|\mathring{U}_i^j|$-th and $(|\mathring{U}_i^j| + 1)$-th nearest neighbor samples to c_i^j, respectively.

Since traditional granular-balls cannot be directly generated in multi-source environments, we propose a neighborhood search-based generation method tailored for these systems. The method employs a neighborhood search strategy, and Fig. 2 shows its flowchart. This approach relies on the homogeneity of local neighborhoods, where data points typically possess one label and thus form a granular-ball.

To address the issue of overlapping granular-balls, a metric termed the overlap ratio is proposed to measure the extent of inter-ball overlap.

Definition 2. For any granular-ball $gb_i^j \in msgb_i$ within the multi-source granular-ball set $MSGB$, its overlap ratio is defined as follows:

$$\mathcal{OR}(gb_i^j) = \frac{\left| \bigcup_{t \neq i} (\mathring{U}_i^j \cap \mathring{U}_t^j) \right|}{\left| \mathring{U}_i^j \right|}. \tag{12}$$

By restricting this overlap ratio, the algorithm effectively avoids the formation of highly intersected granular-balls, ensuring a compact and distinct representation space.

The operational mechanism of the proposed generation method relies on a binary search to identify the maximum number of nearest neighbors around an initial sample. This forms a granular-ball that meets the formation criteria. This process is repeated across all other information sources. Finally, the granular-balls from all sources are combined into a multi-source granular-ball. The detailed generation algorithm is shown in Algorithms 1 and 2.

In the proposed algorithm, the target purity threshold T governs the label consistency within each granular-ball. Setting T to a strictly high value, typically between 0.90 and 0.95, guarantees the homogeneity of the generated balls. Additionally, the size boundary $[\underline{M}, \overline{M}]$ is implemented to regulate the granularity. The upper bound restricts the generation of excessively large balls to preserve local structural details, whereas the lower bound mitigates the noise sensitivity caused by overly small balls. The specific values in our experiments are determined empirically.

Algorithm 1: Multi-Source Granular-Ball Generation Algorithm.

Input: Granular-ball size limit $[\underline{M}, \overline{M}]$, sample x, source index j, target purity T, number of sources q.
Output: A multi-source granular-ball.

$left = \underline{M}$, $right = \overline{M}$, $mid = \lfloor \frac{left+right}{2} \rfloor$, $gb_i^j = \emptyset$, $msgb_i = \emptyset$;
while $left < right$ **do**
  Query the mid nearest neighbor samples $\mathring{U}_i'^j$ of sample x;
  Construct a granular-ball $gb_i'^j = (\mathring{U}_i'^j, c_i'^j, r_i'^j, \mathring{d}_i'^j, p_i'^j)$ via Definition 1;
  if $p_i'^j \geq T$ **then**
    $left = mid + 1$, $mid = \lfloor \frac{left+right}{2} \rfloor$;
    if $gb_i^j = \emptyset$ ***or*** $|gb_i'^j| > |gb_i^j|$ **then**
      $gb_i^j = gb_i'^j$;
    end
  else
    $right = mid - 1$, $mid = \lfloor \frac{left+right}{2} \rfloor$;
  end
end
if $gb_i^j = \emptyset$ ***or*** $\mathcal{OR}(gb_i^j) > 0$ **then**
  return None ;
end
$msgb_i \leftarrow gb_i^j$;
for $j = 1$ **to** q **do**
  Query the mid nearest neighbor samples $\mathring{U}_i'^j$ of c_i^j;
  Construct a granular-ball $gb_i^j = (\mathring{U}_i'^j, c_i'^j, r_i'^j, \mathring{d}_i'^j, p_i'^j)$ via Definition 1;
  if $\mathring{d}_i'^j = \mathring{d}_i^j \subsetneq \mathcal{OR}(gb_i^j) \leq 0$ **then**
    $msgb_i \leftarrow gb_i^j$;
  end
end
return $msgb_i$;

Algorithm 2: Neighbor Search-Based Generation of Multi-Source Granular-Balls.

Input: Dataset U, target purity T, granular-ball size limit $[\underline{M}, \overline{M}]$.
Output: Multi-source granular-ball set $MSGB$.

$S = \emptyset$, $MSGB = \emptyset$;
while $|S| < |U|$ **do**
  Randomly select a sample x from $(U - S)$;
  Obtain a multi-source granular-ball $msgb_i$ via Algorithm 1;
  $S \leftarrow x$;
  Retrieve $gb_i^j = (\mathring{U}_i^j, c_i^j, r_i^j, \mathring{d}_i^j, p_i^j) \in msgb_i$;
  if $msgb_i$ ***is not*** $None$ **then**
    $S = S \cup \mathring{U}_i^j$, $MSGB \leftarrow msgb_i$;
  end
end
return $MSGB$;

4. Feature selection based on multi-source granular-ball rough set and PROMETHEE method

Feature selection is a crucial preprocessing step in data mining and machine learning. It is particularly important in high-dimensional data scenarios. Rough set theory, as an effective tool for data analysis, has been widely applied in feature selection. However, traditional rough set models are primarily designed for single-source information systems. Consequently, they struggle to handle feature selection problems in multi-source information systems directly.

Granular-ball rough set, which combines granular-ball computing with rough set theory, has emerged as an effective approach for handling uncertainty. It constructs granular-ball neighborhoods efficiently, significantly improving the speed and stability of rough set computations. Nevertheless, existing granular-ball rough set methods are also confined to single-source information systems. Their single-perspective framework is fundamentally inadequate for capturing the more complex structural characteristics and potential

inter-source relationships present in multi-source environments. Data in multi-source information systems typically come from multiple homogeneous sources. These sources share the same structure but differ in distribution or quality. This diversity and heterogeneity pose greater challenges for traditional feature selection methods.

To overcome these limitations, this paper proposes a new theoretical framework called the Multi-Source Granular-Ball Variable Precision Rough Set (MSGB-VPRS). This framework extends the traditional granular-ball rough set theory to multi-source scenarios. Based on the granular-ball generation strategy introduced in the previous section, the framework implements the definitions for the granular-ball variable precision rough set within multi-source information systems. Furthermore, this paper integrates the PROMETHEE II method. PROMETHEE II is a robust outranking technique widely used in MCDM. By combining it with the multi-source granular-ball rough set, we propose a novel feature selection algorithm. This algorithm can effectively perform feature selection tasks in multi-source information systems.

First, this paper provides the definitions of the lower and upper approximations within the multi-source granular-ball variable precision rough set.

Definition 3. Let $\{DIS_j = (U, AT \cup D, V, f^j) | j = 1, 2, 3, \dots, q\}$ be a multi-source information system, where $U = \{x_1, x_2, \dots, x_m\}$ denotes the sample universe and $AT = \{a_1, a_2, \dots, a_n\}$ is the set of conditional attributes. Given the multi-source granular-ball set $MSGB$ and a classification threshold $0 \le \beta < 0.5$, the lower and upper approximations of the target subset $X_k \subseteq U$ under the MSGB-VPRS model, with respect to the j-th information source, are defined as follows:

$$\underline{GBR}_{\beta}^{j}(X_k) = \left\{ x \in U \;\middle|\; \frac{|\mathring{U}_i^j \cap X_k|}{|\mathring{U}_i^j|} \ge 1 - \beta \right\}, \tag{13}$$

$$\overline{GBR}_{\beta}^{j}(X_k) = \left\{ x \in U \;\middle|\; \frac{|\mathring{U}_i^j \cap X_k|}{|\mathring{U}_i^j|} > \beta \right\}. \tag{14}$$

Next, to more accurately quantify the uncertainty in multi-source systems, this paper employs Zentropy [26,27] as an uncertainty measure. Zentropy is a novel entropy measure designed for granular structures. Unlike traditional entropy, it provides a multi-level analysis of uncertainty, capturing the dynamics of certainty across hierarchical levels. This capability allows it to capture the trend of certainty changes, from coarse to fine or fine to coarse, within a hierarchical structure.

In this paper, we appropriately extend Zentropy to enable its application in describing uncertainty across different information sources.

Definition 4. Given a multi-source information system $\{DIS_j = (U, AT \cup D, V, f^j) | j = 1, 2, 3, \dots, q\}$, let $A \subseteq AT$ be an attribute subset. The multi-source granular-ball set generated by Algorithm 2 is denoted as $MSGB = \{msgb_1, msgb_2, \dots, msgb_p\}$, where $msgb_i = \{gb_i^1, gb_i^2, \dots, gb_i^q\}$. For the target set $U/D = \{X_1, X_2, \dots, X_s\}$, the Zentropy for the j-th source is defined as follows:

$$Z^j(A, D) = -\sum_{k=1}^{s} p_k \log_2 p_k + \sum_{k=1}^{s} p_k Z_k^j, \tag{15}$$

where $p_k = \frac{|X_k|}{|U|}$ denotes the probability of the k-th decision class, and Z_k^j represents the uncertainty of this class at a finer granularity level.

The uncertainty Z_k^j at the approximation level is defined as:

$$Z_k^j = -\sum_{l=1}^{2} p_{kl} \log_2 p_{kl} + \sum_{l=1}^{2} p_{kl} Z_{kl}^j, \tag{16}$$

where $p_{k1} = \frac{\left|\underline{GBR}_{\beta}^{j}(X_k)\right|}{|X_k|}$ denotes the proportion of samples in class X_k that are deterministically classified by the lower approximation, $p_{k2} = \frac{\left|X_k - \underline{GBR}_{\beta}^{j}(X_k)\right|}{|X_k|}$ denotes the proportion of non-deterministic samples.

The entropy for the uncertainty at the sample level is defined as:

$$Z_{kl}^j = -\sum_{i=1}^{|N_{kl}|} p_{kli} \log_2 p_{kli} + \sum_{i=1}^{|N_{kl}|} p_{kli} Z_{kli}^j, \tag{17}$$

where $N_{k1} = \underline{GBR}_{\beta}^{j}(X_k)$ and $N_{k2} = X_k - \underline{GBR}_{\beta}^{j}(X_k)$. Here, $p_{kli} = \frac{\left|\mathring{U}_i^j\right|}{\sum_{i=1}^{p}\left|\mathring{U}_i^j\right|}$ denotes the probability of the i-th granular-ball neighborhood category within this source.

Finally, the uncertainty at the finest granularity is defined as:

$$Z_{kli}^j = -\sum_{o=1}^{2} p_{klio} \log_2 p_{klio}, \tag{18}$$

where $p_{kli1} = \frac{\left|\mathring{U}_i^j \cap X_k\right|}{|\mathring{U}_i^j|}$ and $p_{kli2} = \frac{\left|\mathring{U}_i^j \cap X_k^c\right|}{|\mathring{U}_i^j|}$.

To achieve feature selection, this paper integrates the PROMETHEE method with the proposed multi-source granular-ball variable precision rough set model. We propose a novel feature selection framework specifically designed for direct implementation in multi-source information systems. The basic procedure is as follows. First, the information gain based on Zentropy is calculated within each information source. Then, the information gains from different sources are treated as multiple criteria. Using the PROMETHEE method, all attributes are comprehensively ranked based on these criteria. Finally, the top k features are selected. The complete algorithm is described below.

Step 1 **Construct Multi-source Granular-balls:** Given the multi-source information system $\{DIS_j = (U, AT \cup D, V, f^j) | j = 1, 2, 3, \ldots, q\}$, use Algorithm 2 to construct the multi-source granular-ball set $MSGB$. This set serves as a unified foundational structure for evaluating attribute uncertainty across different information sources.

Step 2 **Calculate Zentropy-based Information Gain on Different Sources:** For each candidate attribute $a_t \in AT$, compute the change in Zentropy-based uncertainty, i.e., the information gain, on each information source j. The information gain is defined as:

$$IG^j(a_t) = Z^j(AT - \{a_t\}, D) - Z^j(AT, D). \tag{19}$$

A larger information gain indicates a greater contribution of attribute a_t to reducing uncertainty on source j.

Step 3 **Construct the Preference Matrix between Attributes:** The PROMETHEE method uses preference functions to quantify the relative advantage of any two attributes under a given criterion. For attributes a_t and a_u, their difference in information gain on the j-th source is defined as:

$$d_j(a_t, a_u) = IG^j(a_t) - IG^j(a_u). \tag{20}$$

To map this difference to a preference degree, this paper employs the common linear preference function:

$$P_j(d_j) = \begin{cases} 0, & d_j \le 0, \\ \dfrac{d_j}{p_j}, & 0 < d_j < p_j, \\ 1, & d_j \ge p_j, \end{cases} \tag{21}$$

where p_j is the preference threshold for source j, set to $p_j = 1$ in the experiments. The linear preference function is employed to translate differences in information gains into a [0, 1] preference intensity without introducing unwarranted complexity. This approach is standard in PROMETHEE-based feature ranking [22,23], as it directly converts continuous metric differences into preference flows. Thus, the preference degree of attribute a_t over attribute a_u on source j is:

$$\pi_j(a_t, a_u) = P_j\left(IG^j(a_t) - IG^j(a_u)\right). \tag{22}$$

Step 4 **Weighted Aggregation of Preferences from Multi-source Criteria:** The preference degrees from all information sources are combined using weights to obtain a global preference index for each attribute pair:

$$\pi(a_t, a_u) = \sum_{j=1}^{q} w_j \, \pi_j(a_t, a_u), \qquad \sum_{j=1}^{q} w_j = 1. \tag{23}$$

This paper uses equal weights, therefore:

$$\pi(a_t, a_u) = \frac{1}{q} \sum_{j=1}^{q} \pi_j(a_t, a_u). \tag{24}$$

Step 5 **Flow Calculation and Global Ranking:** PROMETHEE ranks the attributes by calculating the positive, negative, and net flows. The positive flow measures the strength of attribute a_t relative to all others:

$$\phi^+(a_t) = \frac{1}{n-1} \sum_{\substack{u=1 \\ u \ne t}}^{n} \pi(a_t, a_u). \tag{25}$$

The negative flow measures the weakness of attribute a_t, i.e., the extent to which it is outperformed by others:

$$\phi^-(a_t) = \frac{1}{n-1} \sum_{\substack{u=1 \\ u \ne t}}^{n} \pi(a_u, a_t). \tag{26}$$

The net flow, used as the final ranking criterion, is defined as:

$$\phi(a_t) = \phi^+(a_t) - \phi^-(a_t). \tag{27}$$

A larger net flow indicates better overall performance of attribute a_t.

Step 6 **Select the Feature Subset:** Rank all attributes in descending order based on their net flow values $\phi(a_t)$. Finally, select the top-k attributes as the feature subset, forming the final feature set with optimal discriminative power.

5. Information fusion

The MSGB-VPRS framework, combined with the PROMETHEE method, can be effectively extended to address information fusion challenges on multi-source homogeneous datasets. The fundamental idea is to utilize the Zentropy measure to evaluate each source, rank them to determine the optimal source for each feature, and finally fuse the optimal sources of all features to achieve effective information fusion. Following a similar stepwise procedure as in Section 4, the information fusion strategy is outlined below.

Step 1 **Construction of Multi-Source Information System:** Let $\mathcal{U}$ denote a common set of objects, and assume there exist m homogeneous information sources $S_1, S_2, \ldots, S_m$, where each source provides a set of conditional attributes $\mathcal{A}_i$ over $\mathcal{U}$ with a shared decision attribute Y.

Step 2 **Computation of Rough Set Entropy:** For each source S_i, a rough set-based uncertainty measure, such as conditional entropy $H(Y|\mathcal{A}_i)$, discernibility entropy, or neighborhood entropy, is calculated to assess the quality or discriminative power of its feature subset. This yields an evaluation matrix $\mathbf{D} = [d_{ij}]$, where d_{ij} represents the j-th entropy measure for the i-th source.

Step 3 **Normalization of the Decision Matrix:** To ensure comparability across different criteria, the decision matrix $\mathbf{D}$ is normalized. For entropy-based criteria, which are minimized, the normalization can be performed as:

$$d_{ij}^{\text{norm}} = \frac{\max_j d_{ij} - d_{ij}}{\max_j d_{ij} - \min_j d_{ij}}. \tag{28}$$

Step 4 **Construction of Preference Functions:** For each pair of sources (S_i, S_j) and each criterion k, the preference function $P_k(S_i, S_j)$ is computed based on the normalized difference in entropy scores. A commonly used preference function is:

$$P_k(S_i, S_j) = \begin{cases} 0, & \text{if } d_{ik}^{\text{norm}} - d_{jk}^{\text{norm}} \le 0, \\ d_{ik}^{\text{norm}} - d_{jk}^{\text{norm}}, & \text{otherwise.} \end{cases} \tag{29}$$

The aggregated preference degree is then obtained by weighting all criteria:

$$\pi(S_i, S_j) = \sum_{k=1}^{n} w_k \cdot P_k(S_i, S_j), \tag{30}$$

where w_k denotes the weight of the k-th criterion.

Step 5 **Computation of Net Flow Scores:** Using the aggregated preferences, the positive flow $\phi^+(S_i)$, negative flow $\phi^-(S_i)$, and net flow $\phi(S_i)$ are computed as follows:

$$\phi^+(S_i) = \frac{1}{m-1} \sum_{j \neq i} \pi(S_i, S_j), \tag{31}$$

$$\phi^-(S_i) = \frac{1}{m-1} \sum_{j \neq i} \pi(S_j, S_i), \tag{32}$$

$$\phi(S_i) = \phi^+(S_i) - \phi^-(S_i). \tag{33}$$

Step 6 **Ranking and Fusion:** The information sources are ranked according to their net flow scores. The top-ranked sources are selected for information fusion, which can be performed through feature concatenation, weighted combination, or other integration techniques, depending on the downstream application requirements.

This framework provides a principled and interpretable strategy to fuse multi-source data by leveraging rough set entropy for uncertainty evaluation and PROMETHEE for source-level ranking and selection.

6. Experimental results and analysis

This section conducts a comprehensive performance evaluation of the proposed multi-source granular-ball feature selection method (MSGBP) and information fusion strategy (MSGB-IFS). The experiments employ three classic classifiers (Gradient Boosting [28], SVM [29] and XGBoost [30]). The proposed method is compared against the raw feature set (Raw) and mainstream methods including GBRS [13], GBNRS [31] and MEL [32], as well as RST-P [22] and FSPA-MODI [23], both of which are PROMETHEE-based. Additionally, the performance of the proposed fusion strategy is compared against several other information fusion strategies (MAX_FUSION, MEAN_FUSION, PCA_FUSION). For each algorithm, the average test accuracy across different datasets is calculated to evaluate its classification performance, and the standard deviation is computed to assess the stability of the results. The experiments utilize 12 classic datasets from the UCI machine learning repository, whose detailed information is presented in Table 1.

To ensure all features are compared on the same scale, each dataset undergoes a normalization preprocessing step. For any feature x_j, the min-max normalization method is employed to scale its values to the interval $[0, 1]$, calculated as follows:

$$x'_j = \frac{x_j - \min(x_j)}{\max(x_j) - \min(x_j)}$$

where $\min(x_j)$ and $\max(x_j)$ represent the minimum and maximum values of feature x_j across the entire dataset, respectively.

Table 1
Basic information of datasets.

No.	Dataset	Samples	Features	Classes	No.	Dataset	Samples	Features	Classes
1	Cancer	116	9	2	7	Letter	20,000	16	26
2	German	1000	24	2	8	Mushroom	7535	22	2
3	Grid	10,000	13	2	9	Qsar	1055	41	2
4	Heart1	294	13	2	10	Quality	4898	11	7
5	Htru3	8011	8	2	11	Spambase	4601	57	2
6	Ionosphere	351	34	2	12	Urban	675	147	9

The datasets used in the experiments are sourced from the UCI Machine Learning Repository. To simulate a multi-source information environment, multiple homogeneous but slightly varied data sources are generated from each original dataset. This is done by adding Gaussian white noise with a mean of 0 and a variance of $\sigma^2 = 0.1$ to the normalized original data matrix. The detailed construction method is described in Section 6.

Table 2
Classification accuracy (% ± std) comparison of Gradient Boosting with different methods.

Dataset	Raw	GBRS	GBNRS	MEL	RST-P	FSPA-MODI	MSGBP
Cancer	74.39±13.3	74.32±10.0	75.15±11.4	68.52±16.7	66.44±12.6	67.27±15.0	**75.32±13.3**
German	75.70±2.36	70.10±3.35	75.80±2.20	74.00±2.49	69.40±3.63	67.50±1.72	**76.50±3.26**
Grid	90.16±1.49	**91.62±0.81**	90.30±1.42	83.13±3.10	77.69±1.43	82.95±1.30	87.94±2.74
Heart1	71.08±2.96	73.46±3.58	71.08±2.96	73.60±6.17	**74.47±5.21**	72.87±8.87	74.14±6.95
Htru3	97.29±0.66	97.22±0.61	**97.38±0.50**	96.45±0.55	97.19±0.57	97.25±0.61	97.34±0.52
Ionosphere	92.60±1.98	93.46±3.53	93.16±3.86	89.74±4.94	94.59±2.50	92.57±5.08	**97.22±0.00**
Letter	**88.66±1.07**	75.33±0.75	76.80±4.12	87.20±0.80	57.10±1.11	80.69±1.16	85.06±2.50
Mushroom	99.96±0.06	99.92±0.09	99.93±0.09	99.41±0.37	99.31±0.26	99.91±0.13	**99.99±0.04**
Qsar	85.22±2.86	77.25±1.92	84.94±2.16	84.57±2.53	84.64±1.76	82.56±4.23	**86.62±2.25**
Quality	55.88±2.75	51.69±1.99	55.88±2.75	54.08±1.95	48.96±1.29	48.45±1.60	**56.07±2.56**
Spambase	94.33±0.91	87.55±1.56	94.52±0.97	92.44±1.34	94.28±0.84	91.96±0.77	**95.07±1.05**
Urban	85.35±4.55	85.35±4.11	85.65±4.63	83.78±4.29	78.37±2.99	84.31±5.15	**85.75±3.77**
Average	84.22±2.91	81.44±2.69	83.38±3.09	82.24±3.77	78.54±2.85	80.69±3.80	**84.75±3.25**

Since the selected UCI datasets are inherently single-source, we construct synthetic multi-source datasets by adding Gaussian white noise to simulate a multi-source information environment. Specifically, based on the normalized original data matrix X, the i-th synthetic data source S_i is generated as:

$$S_i = X + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

where ϵ_i is a noise matrix whose elements are independent and identically distributed, following a normal distribution with a mean of 0 and a variance of $\sigma^2 = 0.1$. This approach allows the generation of multiple homogeneous yet slightly varied data sources from each original dataset, thereby facilitating the validation of the algorithm's effectiveness in multi-source scenarios.

In the experiments, key parameters are configured based on established practices and empirical tuning. For the multi-source granular-ball generation, the target purity threshold T is set to 0.95, and the size boundary $[\underline{M}, \overline{M}]$ is constrained to [5, 50] to balance local structural preservation with noise suppression. In the feature selection framework, the preference threshold p_j for the PROMETHEE linear preference function is uniformly set to 1.0 across all sources. This setting strictly follows the linear preference function defined in Eq. (21). For the VPRS model, the classification error threshold β is empirically set to 0.1 to maintain robustness against misclassification while preserving approximation accuracy. Additionally, the number of sources q is fixed at 3 to emulate a representative multi-source environment without incurring excessive computational overhead.

6.1. Performance comparison of feature selection methods

Tables 2–4 present the classification accuracy of different feature selection algorithms across three classifiers. The results demonstrate the superior performance of our proposed method on multiple classifiers. Specifically, across Gradient Boosting, SVM and XGBoost, MSGBP consistently matches or outperforms the baselines on most datasets. The average accuracies reach 84.75%, 82.05%, and 85.50%, respectively, which are the highest among all methods. Compared to using the raw feature set (Raw) and other mainstream feature selection algorithms (GBRS, GBNRS, MEL, RST-P, and FSPA-MODI), the improvement in accuracy achieved by the MSGBP method is significant. These results validate that the MSGBP method can effectively select discriminative features in multi-source environments, thereby enhancing the generalization capability of classification models.

Based on the accuracy data from Tables 2–4, the results are converted into rank orders, and the corresponding three-dimensional histograms are visualized in Fig. 3. A lower rank value indicates better performance and a higher position. The figure intuitively shows that MSGBP method consistently achieves high rankings across all classifiers.

Table 3
Classification accuracy (% ± std) comparison of SVM with different methods.

Dataset	Raw	GBRS	GBNRS	MEL	RST-P	FSPA-MODI	MSGBP
Cancer	66.52±10.4	63.94±8.89	68.18±12.7	62.52±10.9	69.77±14.8	60.38±7.77	**71.00±11.9**
German	76.90±3.18	70.00±0.00	76.90±3.03	73.35±2.06	70.00±0.00	70.00±0.00	**77.30±2.75**
Grid	85.96±1.30	81.65±1.79	**85.97±1.28**	82.47±2.06	73.59±1.48	82.93±1.20	84.69±1.44
Heart1	74.21±5.46	75.44±5.82	74.53±5.10	74.94±6.34	78.18±5.76	73.53±5.59	**78.53±5.59**
Htru3	96.88±0.82	96.84±0.83	96.83±0.80	96.38±0.74	96.87±0.83	96.78±0.85	96.86±0.80
Ionosphere	87.45±5.44	87.47±3.84	88.87±5.14	83.41±6.05	88.03±4.82	87.75±5.71	**92.36±2.66**
Letter	**82.09±0.66**	61.12±1.01	70.70±1.84	74.92±0.72	41.52±0.69	67.10±1.33	75.01±3.40
Mushroom	**95.22±1.15**	95.01±1.15	94.53±1.18	73.62±2.13	89.22±1.40	89.56±1.29	94.96±1.16
Qsar	85.22±1.69	67.58±0.96	85.69±2.49	81.65±2.64	84.46±1.74	80.19±3.60	**86.24±2.35**
Quality	51.96±1.63	45.47±0.38	51.96±1.63	51.90±1.50	44.88±0.05	44.88±0.05	**51.98±1.55**
Spambase	90.09±0.88	75.57±1.53	**90.22±0.94**	84.47±1.68	89.74±0.98	83.87±1.29	89.89±0.79
Urban	85.79±3.79	**85.94±3.98**	85.49±3.77	84.90±3.78	76.75±3.90	83.56±2.86	85.78±3.33
Average	81.52±3.04	75.50±2.52	80.82±3.32	77.04±3.38	75.25±3.04	76.71±2.63	**82.05±3.15**

Table 4
Classification accuracy (% ± std) comparison of XGBoost with different methods.

Dataset	Raw	GBRS	GBNRS	MEL	RST-P	FSPA-MODI	MSGBP
Cancer	75.23±14.9	75.91±15.6	**80.30±13.1**	70.86±14.0	71.67±14.1	69.09±15.6	73.65±14.9
German	75.40±3.81	66.80±3.49	76.30±3.74	73.60±2.52	68.60±5.08	66.40±4.55	**77.00±3.30**
Grid	93.90±0.82	**94.34±0.51**	93.99±0.87	82.41±3.72	76.39±1.75	81.58±1.19	88.50±3.89
Heart1	68.33±5.75	71.39±5.50	70.06±4.79	71.19±6.76	71.08±7.43	71.14±8.25	**72.37±4.08**
Htru3	97.42±0.65	97.40±0.76	97.44±0.58	96.92±0.58	97.38±0.71	97.24±0.61	**97.48±0.57**
Ionosphere	91.75±3.65	92.03±3.48	91.75±2.80	89.85±5.05	93.72±3.25	92.02±4.44	**97.22±0.00**
Letter	**94.50±0.32**	83.20±0.71	80.55±3.34	92.91±0.43	65.40±0.51	87.66±0.75	91.37±1.86
Mushroom	99.93±0.07	99.96±0.06	**99.99±0.04**	99.30±0.42	99.50±0.23	99.95±0.09	99.98±0.05
Qsar	84.84±2.52	75.73±4.09	86.17±2.01	84.59±2.50	85.40±2.24	82.75±3.78	**86.71±2.57**
Quality	58.64±1.91	55.08±1.26	58.64±1.91	53.94±2.47	49.16±2.22	49.69±1.11	**59.25±2.04**
Spambase	94.57±0.55	87.44±1.41	94.83±0.63	92.07±1.38	94.78±0.45	91.57±1.07	**95.02±1.06**
Urban	87.41±4.29	87.71±5.29	**87.71±4.60**	84.97±3.83	81.33±2.95	83.71±4.06	87.40±3.59
Average	85.16±3.27	82.25±3.52	84.81±3.20	82.72±3.64	79.53±3.41	81.07±3.79	**85.50±3.16**

Subsequently, the Friedman test is conducted to assess the statistical significance of the observed differences in performance. The Friedman test statistic t_{χ^2} is calculated as follows:

$$t_{\chi^2} = \frac{12n}{g(g+1)}\left(\sum_{i=1}^{g} R_i^2 - \frac{g(g+1)^2}{4}\right) \tag{34}$$

where n denotes the number of datasets (here, $n = 12$), g denotes the number of methods compared (here, $g = 7$), and R_i represents the average rank of the i-th method.

Based on the average rankings of each method under each classifier, the calculated Friedman test statistics t_{χ^2} are: Gradient Boosting: 33.94, SVM: 31.49, and XGBoost: 26.74. The adjusted Friedman statistic t_F is then computed as:

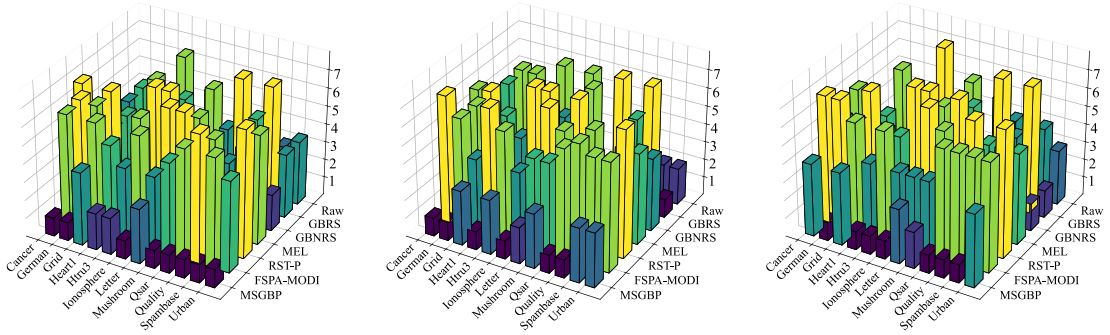
$$t_F = \frac{(n-1)t_{\chi^2}}{n(g-1) - t_{\chi^2}} \tag{35}$$

The corresponding t_F values are 9.81, 8.55, and 6.50, respectively. At a significance level of $\alpha = 0.05$, all three t_F values far exceed the critical value of the F-distribution, $F(6, 66) \approx 2.24$. This indicates that there are statistically significant differences in the performance rankings among the different feature selection methods for all three classifiers. According to the average rankings and Friedman test results summarized in Table 5, the p-values for all three classifiers are well below the 0.05 significance threshold. This confirms that statistically significant performance differences exist among the seven evaluated methods. A subsequent Nemenyi post-hoc test was therefore conducted to delineate the specific pairwise differences between the algorithms.

The Critical Difference (CD) is calculated using the formula:

$$CD = q_\alpha \sqrt{\frac{g(g+1)}{6n}} \tag{36}$$

where q_α is the critical value (for $\alpha = 0.05$, $g = 7$ and $q_\alpha = 2.949$). Substituting the values yields $CD = 2.60$. The CD value serves as the threshold for significance. If the difference between the average ranks of any two methods is greater than the CD, we can reject the null hypothesis that their performances are equivalent at the $\alpha = 0.05$ significance level. According to the rankings in Table 5, MSGBP consistently secures the top position across all evaluated classifiers. When applying the CD threshold, our method demonstrates statistically significant superiority over the majority of baselines, including MEL, RST-P, and FSPA-MODI in every scenario. It further establishes a significant lead over GBRS when using Gradient Boosting and SVM. While the performance gap between MSGBP and the

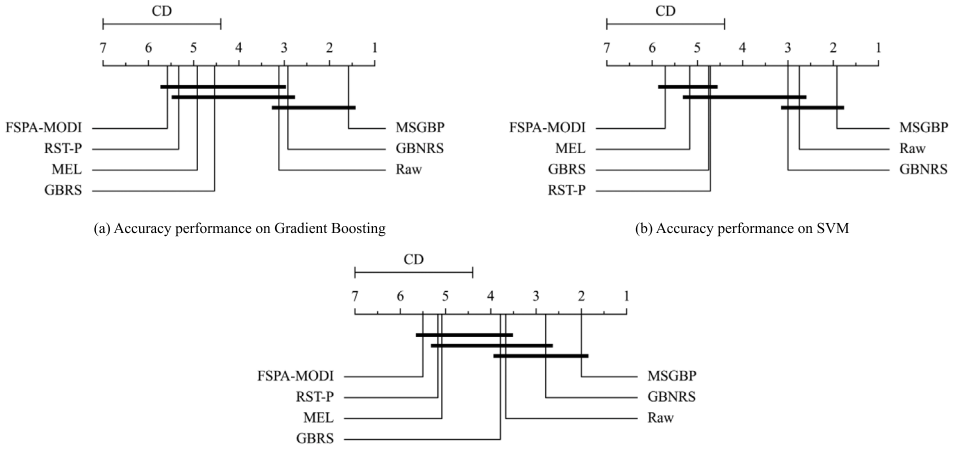


(a) Gradient Boosting (b) SVM (c) XGBoost

Fig. 3. Average ranks of different feature selection methods across the Gradient Boosting, SVM and XGBoost classifiers.

Table 5
Average rankings of different feature selection methods across classifiers.

Classifier	Raw	GBRS	GBNRS	MEL	RST-P	FSPA-MODI	MSGBP	P-Value	CD
Gradient Boosting	3.12	4.54	2.92	4.92	5.33	5.58	1.58	6.92×10^{-6}	2.60
SVM	2.75	4.75	3.00	5.17	4.71	5.71	1.92	2.04×10^{-5}	2.60
XGBoost	3.67	3.79	2.79	5.08	5.17	5.50	2.00	1.62×10^{-4}	2.60



(a) Accuracy performance on Gradient Boosting (b) Accuracy performance on SVM

(c) Accuracy performance on XGBoost

Fig. 4. Results of the significance comparison based on the Nemenyi post-hoc test.

Table 6
Runtime (s) comparison with different methods.

No.	Dataset	RST-P	FSPA-MODI	MSGBP	No.	Dataset	RST-P	FSPA-MODI	MSGBP
1	Cancer	**0.066**	0.345	1.184	7	Letter	3511	76,828	**2949**
2	German	37.92	54.23	**30.31**	8	Mushroom	282.4	2746	**166.2**
3	Grid	705.7	2972	**664.5**	9	Qsar	89.26	127.2	**50.34**
4	Heart1	**0.415**	2.674	2.751	10	Quality	148.6	1451	**121.7**
5	Htru3	332.8	1268	**273.7**	11	Spambase	197.9	3409	**162.4**
6	Ionosphere	15.98	**9.181**	11.27	12	Urban	**48.82**	203.5	153.3

remaining methods falls within the CD margin, MSGBP still maintains a clear advantage. Fig. 4 displays the outcomes of the Nemenyi analysis.

To further validate the computational efficiency advantage of the proposed MSGBP method, we design an efficiency comparison experiment. This experiment aims to control variables and specifically evaluate the efficiency improvement brought by multi-source granular-ball computing to the PROMETHEE-based feature selection algorithm. Therefore, we select two typical baseline algorithms, RST-P and FSPA-MODI, which also integrate the PROMETHEE II method for feature ranking, for comparison. This controlled setup isolates the computational acceleration contributed by the granular-ball structure, thereby accurately highlighting the unique advantages of the method within its specific algorithmic domain.

Table 7
Classification accuracy (% ± std) comparison of Gradient Boosting, SVM and XGBoost with different fusion methods.

Dataset	Gradient Boosting				SVM				XGBoost			
	MAX_FUSION	MEAN_FUSION	PCA_FUSION	MSGB-IFS	MAX_FUSION	MEAN_FUSION	PCA_FUSION	MSGB-IFS	MAX_FUSION	MEAN_FUSION	PCA_FUSION	MSGB-IFS
Cancer	72.65±11.9	74.39±13.3	62.35±17.7	**78.23±13.4**	66.52±10.4	66.52±10.4	**73.33±14.9**	68.82±10.0	71.74±15.0	75.23±14.9	65.76±14.5	**75.52±16.3**
German	74.70±3.06	75.70±2.36	72.20±2.62	**80.70±3.30**	77.70±3.47	76.90±3.18	75.10±3.41	**81.80±3.61**	74.80±2.94	75.40±3.81	71.30±2.11	**78.50±3.41**
Grid	90.66±0.88	90.16±1.49	89.95±1.00	**91.80±2.91**	82.79±1.79	**85.96±1.30**	85.85±1.51	85.80±4.42	93.01±0.81	**93.81±0.64**	91.61±0.81	93.10±2.06
Heart1	72.48±5.64	71.08±2.96	72.48±5.73	**75.42±4.97**	73.54±6.37	74.21±5.46	74.85±3.79	**75.22±5.37**	71.77±6.40	69.34±5.54	73.52±5.49	**74.99±6.14**
Htru3	97.05±0.75	97.29±0.66	97.05±0.78	**97.50±1.86**	96.92±0.82	96.88±0.82	**97.29±0.74**	95.88±1.19	97.48±0.61	97.42±0.65	97.42±0.69	**97.63±2.08**
Ionosphere	92.87±3.63	92.60±1.98	93.14±5.08	**93.37±3.06**	87.44±5.92	87.45±5.44	86.02±4.39	**88.02±4.74**	**93.73±2.63**	91.17±3.40	91.73±4.15	93.15±4.03
Letter	88.69±0.73	88.66±1.07	86.04±0.93	**91.95±3.01**	82.03±0.67	82.09±0.66	84.82±0.80	**85.35±1.93**	94.35±0.43	94.50±0.32	92.10±0.62	**94.90±1.77**
Mushroom	99.99±0.04	99.96±0.06	99.75±0.22	**100.00±0.00**	94.84±1.18	95.22±1.15	**96.03±0.99**	94.49±1.16	99.97±0.06	99.93±0.07	**99.99±0.04**	99.97±0.05
Qsar	85.22±2.26	85.22±2.86	85.02±2.93	**85.81±2.60**	85.22±2.30	85.22±1.69	**86.92±1.09**	85.43±2.19	85.12±2.86	85.97±2.68	84.83±2.05	**86.38±3.04**
Quality	55.12±2.24	55.88±2.75	51.37±1.41	**56.13±1.54**	51.29±1.72	51.96±1.63	51.31±1.52	**54.83±1.62**	56.80±1.97	**58.64±1.91**	51.18±2.41	56.81±2.37
Spambase	**94.46±1.12**	94.33±0.91	91.81±1.03	93.97±0.99	89.83±0.91	90.09±0.88	**92.24±1.08**	89.69±0.87	94.48±1.02	**94.68±0.63**	92.55±0.81	94.21±0.89
Urban	83.57±4.66	85.35±4.55	75.71±4.97	**85.90±2.32**	84.75±3.64	85.79±3.79	84.15±5.98	**86.00±1.72**	86.96±3.59	**87.41±4.29**	78.22±5.16	85.81±4.03
Average	83.96±3.07	84.22±2.92	81.41±3.70	**85.90±3.33**	81.07±3.27	81.52±3.04	82.33±3.35	**82.61±3.24**	85.02±3.20	85.29±3.24	82.52±3.23	**85.91±3.85**

The results are presented in Table 6, which shows that the running time of MSGBP is lower than that of both RST-P and FSPA-MODI on most datasets. This efficiency gain stems from the core innovation of MSGBP, which is to replace raw data points with granular-balls as the fundamental computational units. As described in Section 3, although generating granular-balls incurs an initial cost, once constructed, the universe for subsequent approximation calculations, uncertainty measurement, and preference comparisons is dramatically reduced in scale from the original sample size to the much smaller number of granular-balls. This reduction accelerates the most computationally intensive steps in the PROMETHEE method, namely the calculation of global preference indices and net flows. Therefore, only on a small number of datasets with limited samples does the initialization phase account for a larger fraction of the total time, resulting in slightly longer runtimes compared to the baseline methods.

6.2. Performance comparison of information fusion methods

In the performance comparison of information fusion methods, three commonly used strategies: MAX_FUSION, MEAN_FUSION and PCA_FUSION are evaluated, each based on a distinct fusion principle. MAX_FUSION selects the maximum value from each feature source to highlight the most discriminative features. MEAN_FUSION computes the average value to balance the contributions from multiple sources and enhance stability. PCA_FUSION employs Principal Component Analysis for dimensionality reduction and decorrelation, aiming to retain the most critical information.

Data from Table 7 indicate that the proposed MSGB-IFS method generally outperforms these traditional fusion strategies on most datasets. With the Gradient Boosting classifier, MSGB-IFS achieves the best performance on 11 out of the 12 datasets, attaining an average accuracy of 85.90%, which is higher than those of MAX_FUSION (83.96%), MEAN_FUSION (84.22%), and PCA_FUSION (81.41%). When using the SVM classifier, MSGB-IFS achieves an average accuracy of 82.61%, demonstrating overall superiority over the other methods. Similarly, with the XGBoost classifier, MSGB-IFS also ranks first with an average accuracy of 85.91%. The experimental results confirm that MSGB-IFS can effectively integrate multi-source features and exhibits excellent generalization capability and stability across different classifiers.

7. Conclusion

This paper proposes a novel framework for multi-source information systems, integrating multi-source granular-balls, variable precision rough sets, and the PROMETHEE multi-criteria decision-making method to enhance the performance of feature selection and information fusion. The main strengths of this approach lie in its ability to uniformly represent multi-source data while effectively mitigating noise through granular-balls, robustly measure multi-granular uncertainty using Zentropy, and achieve systematic feature selection via PROMETHEE II preference flow ranking. Experimental results demonstrate that the proposed MSGBP feature selection method and the MSGB-IFS information fusion strategy exhibit superior performance across multiple classifiers and datasets, significantly improving classification accuracy and interpretability. The framework offers significant potential for applications involving noisy, multi-source data, including medical diagnostics and industrial fault prediction. It effectively addresses the challenges of data distribution shifts and environmental noise by selecting robust features and enabling reliable fusion.

Despite these promising results, certain limitations present valuable avenues for future research. First, the existing method assumes that multi-source data share a homogeneous feature space; future work could extend it to heterogeneous multi-source information systems, where different data sources may possess different feature sets or data structures. Second, the current granular-ball generation primarily relies on neighborhood search; adaptive optimization algorithms could be introduced to dynamically adjust the granular-ball purity threshold and radius parameters, thereby enhancing the model's adaptability to complex data distributions. Finally, while the present framework evaluates static datasets, its architecture holds considerable promise for dynamic environments. Because granular-balls localize data representation, the method is intrinsically well-suited for incremental learning scenarios. Future work will focus on developing incremental update mechanisms for multi-source granular-balls, enabling the system to dynamically absorb streaming data and update local structures without the computational overhead of retraining from scratch.

CRediT authorship contribution statement

Lingwei Wei: Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Data curation. **Weirui Ye:** Software, Investigation, Data curation, Conceptualization. **Weihua Xu:** Supervision, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization. **Shuyin Xia:** Methodology, Investigation, Data curation.

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.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 62376229 and in part by the Natural Science Foundation of Chongqing under Grant CSTB2023NSCQ-LZX0027.

Data availability

The link to all the data used for the research described has been shared in the article.

References

[1] K. Cukier, V. Mayer-Schoenberger, The rise of big data: how it's changing the way we think about the world, The Best Writing on Mathematics 2014 (2014) 20–32.
[2] S. Maldonado, J. Pérez, C. Bravo, Cost-based feature selection for support vector machines: an application in credit scoring, Eur. J. Oper. Res. 261 (2) (2017) 656–665.
[3] B. Omar, F. Rustam, A. Mehmood, G.S. Choi, et al., Minimizing the overlapping degree to improve class-imbalanced learning under sparse feature selection: application to fraud detection, IEEE Access 9 (2021) 28101–28110.
[4] X. Li, J. Zhang, F. Safara, Improving the accuracy of diabetes diagnosis applications through a hybrid feature selection algorithm, Neural Process. Lett. 55 (1) (2023) 153–169.
[5] X. Liu, S. Wang, S. Lu, Z. Yin, X. Li, L. Yin, J. Tian, W. Zheng, Adapting feature selection algorithms for the classification of Chinese texts, Systems 11 (9) (2023) 483.
[6] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, J. Mach. Learn. Res. 3 (3) (2003) 1157–1182.
[7] V. Bolón-Canedo, N. Sánchez-Maroño, A. Alonso-Betanzos, Feature selection for high-dimensional data, Prog. Artif. Intell. 5 (2) (2016) 65–75.
[8] A. Jović, K. Brkić, N. Bogunović, A review of feature selection methods with applications, in: 2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), IEEE, 2015, pp. 1200–1205.
[9] Z. Pawlak, Rough sets, International Journal of Computer & Information Sciences 11 (5) (1982) 341–356.
[10] C. Wang, M. Shao, Q. He, Y. Qian, Y. Qi, Feature subset selection based on fuzzy neighborhood rough sets, Knowl.-based Syst. 111 (2016) 173–179.
[11] Z. Qiu, H. Zhao, A fuzzy rough set approach to hierarchical feature selection based on hausdorff distance, Appl. Intell. 52 (10) (2022) 11089–11102.
[12] X. Cao, X. Wang, H. Liang, B. Wei, X. Yang, Open continual sampling with hypersphere knowledge transfer for rapid feature selection, Appl. Soft Comput. 170 (2025) 112664.
[13] S. Xia, C. Wang, G. Wang, X. Gao, W. Ding, J. Yu, Y. Zhai, Z. Chen, GBRS: a unified granular-ball learning model of pawlak rough set and neighborhood rough set, IEEE Trans. Neural Netw. Learn. Syst. 36 (1) (2025) 1719–1733.
[14] S. Cheng, X. Su, B. Chen, H. Chen, D. Peng, Z. Yuan, GBMOD: a granular-ball mean-shift outlier detector, Pattern Recognit. 159 (2025) 111115.
[15] P. Zhang, T. Li, G. Wang, D. Wang, P. Lai, F. Zhang, A multi-source information fusion model for outlier detection, Inf. Fusion 93 (2023) 192–208.
[16] X. Zhang, X. Shen, Graph-driven feature selection via granular-rectangular neighborhood rough sets for interval-valued data sets, Appl. Soft Comput. 170 (2025) 112716.
[17] Z. Feng, X. Zhang, Supervised incremental feature selection using regularization vector for dynamic multi-scale interval valued datasets, Pattern Recognit. 170 (2026) 111985.
[18] W. Li, L. Wei, W. Pedrycz, W. Ding, C. Zhang, T. Zhan, S. Xia, Granular-ball regeneration clustering with principle of justifiable granularity, IEEE Trans. Neural Netw. Learn. Syst. 36 (10) (2025) 18173–18187.
[19] M. Behzadian, R.B. Kazemzadeh, A. Albadvi, M. Aghdasi, PROMETHEE: a comprehensive literature review on methodologies and applications, Eur. J. Oper. Res. 200 (1) (2010) 198–215.
[20] P. Ghasemi, A. Mehdiabadi, C. Spulbar, R. Birau, Ranking of sustainable medical tourism destinations in Iran: an integrated approach using fuzzy SWARA-PROMETHEE, Sustainability 13 (2) (2021) 683.
[21] S.A. Ahmadi, A. Peivandizadeh, Sustainable portfolio optimization model using promethee ranking: a case study of palm oil buyer companies, Discrete Dyn. Nat. Soc. 2022 (1) (2022) 8935213.
[22] W. Xu, Y. Li, Enhancing information fusion and feature selection efficiency via the PROMETHEE method for multi-source dynamic decision data sets, Knowl.-based Syst. 309 (2025) 112781.
[23] W. Xu, Z. Yang, Preference ranking organization method for enrichment evaluation-based feature selection for multiple source ordered information systems, Eng. Appl. Artif. Intell. 142 (2025) 109935.
[24] S. Xia, Y. Liu, X. Ding, G. Wang, H. Yu, Y. Luo, Granular ball computing classifiers for efficient, scalable and robust learning, Inf. Sci. 483 (2019) 136–152.
[25] W. Xu, Z. Tian, Feature selection and information fusion based on preference ranking organization method in interval-valued multi-source decision-making information systems, Inf. Sci. 700 (2025) 121860.
[26] K. Yuan, D. Miao, Y. Yao, H. Zhang, X. Zhao, Feature selection using zentropy-based uncertainty measure, IEEE Trans. Fuzzy Syst. 32 (4) (2024) 2246–2260.
[27] K. Yuan, D. Miao, W. Pedrycz, W. Ding, H. Zhang, Ze-HFS: zentropy-based uncertainty measure for heterogeneous feature selection and knowledge discovery, IEEE Trans. Knowl. Data Eng. 36 (11) (2024) 7326–7339.
[28] J.H. Friedman, Greedy function approximation: a gradient boosting machine, Ann. Stat. 29 (5) (2001) 1189–1232.
[29] J.C. Platt, Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods, in: Advances in Large Margin Classifiers, MIT Press, 1999, pp. 61–74.
[30] T. Chen, C. Guestrin, XGBoost: a scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794.
[31] S. Xia, H. Zhang, W. Li, G. Wang, E. Giem, Z. Chen, GBNRS: a novel rough set algorithm for fast adaptive attribute reduction in classification, IEEE Trans. Knowl. Data Eng. 34 (3) (2022) 1231–1242.
[32] X. Wang, H. Shangguan, F. Huang, S. Wu, W. Jia, MEL: efficient multi-task evolutionary learning for high-dimensional feature selection, IEEE Trans. Knowl. Data Eng. 36 (8) (2024) 4020–4033.