Engineering Applications of Artificial Intelligence 161 (2025) 112007

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

Maximum relevant minimum redundant multi-label feature selection using ant colony optimization

Mohammad Hatami [a], Parham Moradi [a,b,*], Sadegh Sulaimany [a], Mahdi Jalili [b]

[a] Department of Computer Engineering, University of Kurdistan, Sanandaj, Iran
[b] School of Electrical and Electronic Engineering, RMIT University, Melbourne, Australia

ARTICLE INFO

Keywords:
Feature selection
Multi-label data
Mutual information
Graph clustering
Ant colony optimization

ABSTRACT

Multi-label learning tasks involve instances that may belong to multiple categories simultaneously, making feature selection particularly challenging in high-dimensional feature spaces. Existing multi-label feature selection methods often suffer from limitations such as high computational complexity, inadequate handling of feature redundancy, and insufficient modelling of label dependencies. To overcome these challenges, we propose a novel framework called Maximum Relevant Minimum Redundant Multi-Label Feature Selection (MR2MLFS), which integrates a two-layer graph representation with a modified Ant Colony Optimization (ACO) strategy. The first graph layer clusters correlated features using Louvain community detection, while the second constructs a meta-graph to model inter-cluster relationships. ACO then explores this structure, favouring the selection of highly relevant and non-redundant features. To reduce computational overhead, we introduce an information-theoretic metric that estimates both feature-label relevance and feature-feature redundancy, eliminating the need for repeated classifier training during the search. We evaluated the proposed method on ten benchmark multi-label datasets using several multi-label classifiers. Experimental results show that the proposed method outperforms six state-of-the-art methods across multiple evaluation metrics, achieving an average relative improvement of 5–12 % while reducing feature dimensionality by up to 80 %. These results confirm the method's robustness, efficiency, and effectiveness in multi-label feature selection.

1. Introduction

With the rapid growth of data from social networks and digital platforms, reducing data dimensionality has become essential for improving machine learning performance. Feature selection (FS) addresses this by eliminating irrelevant and redundant features while retaining the most informative ones. Traditional FS methods often rely on classifiers (wrappers), which are computationally expensive, whereas filter methods offer greater efficiency by using information-theoretic metrics. While early FS techniques focused on single-label tasks, many real-world problems such as image annotation, document categorisation, and bioinformatics are inherently multi-label. In these tasks, each instance may be associated with multiple class labels, often with complex inter-label correlations (e.g., "flower" is more likely to co-occur with "tree" than with "ship").

Multi-label FS methods are generally categorised as problem transformation or adaptive approaches. The former convert multi-label problems into multiple single-label tasks, using methods like Binary Relevance and Label Power Set. However, these approaches may overlook label dependencies, reducing their effectiveness. Adaptive methods directly explore the feature space without transforming the problem. Notable examples include MICO (Sun et al., 2019), which uses constrained convex optimization; LFD (Lee et al., 2019), based on label frequency difference; MDFS (Zhang et al., 2019a), which incorporates manifold regularization; and MGFS (Hashemi et al., 2020b), which applies PageRank centrality. While effective, most rely on heuristic searches and often lack mechanisms to escape local optima or to jointly consider feature relevance and redundancy. To address these limitations, evolutionary and swarm intelligence-based methods, particularly Ant Colony Optimization (ACO), have shown promise. ACO, as a multi-agent system with strong exploration and exploitation capabilities, has been successfully applied in multi-label FS tasks (Hashemi et al., 2020b). For instance, MLACO (Paniri et al., 2019) uses ACO to select feature subsets based on heuristic evaluations, but it does not

* Corresponding author. Department of Computer Engineering, University of Kurdistan, Sanandaj, Iran.
E-mail addresses: m.hatami@uok.ac.ir (M. Hatami), parham.moradi@rmit.edu.au (P. Moradi), S.Sulaimany@uok.ac.ir (S. Sulaimany), mahdi.jalili@rmit.edu.au (M. Jalili).

https://doi.org/10.1016/j.engappai.2025.112007
Received 18 March 2025; Received in revised form 5 July 2025; Accepted 10 August 2025
Available online 1 September 2025

explicitly address feature redundancy or relevance during the search process.

While multi-label feature selection has gained increasing attention due to the growing presence of high-dimensional multi-label datasets in domains such as bioinformatics, image annotation, and text classification, existing methods still face several practical limitations. Many approaches rely heavily on logical label representations, which assume that all labels associated with an instance are equally informative. However, in real-world scenarios, the significance of each label often varies across instances, leading to suboptimal supervision when these differences are ignored. Another limitation is the prevalent assumption of label independence. Most traditional methods treat labels as unrelated entities, disregarding the semantic and statistical correlations that naturally exist among labels. This simplification can result in the selection of features that are only marginally relevant or fail to capture the complex inter-label interactions critical for performance. Furthermore, many existing algorithms assume complete data availability. In practice, missing features or incomplete label assignments are common in large-scale and heterogeneous datasets. Current methods often struggle to handle such incompleteness effectively, which significantly limits their scalability and applicability to real-world data. Finally, computational complexity and scalability remain critical challenges. While some methods offer high accuracy, they often do so at the cost of increased time and resource requirements, making them impractical for large datasets without additional optimization techniques such as parallelization or dimensionality pre-filtering. By recognizing and explicitly addressing these limitations, our proposed method aims to provide a more robust, flexible, and scalable solution for multi-label feature selection that better reflects the complexities of real-world data environments.

Previous studies, such as those proposed in (Moradi and Rostami, 2015; Tabakhi and Moradi, 2015), have shown that the ACO outperforms other metaheuristic approaches, including Genetic Algorithms (GA) and Particle Swarm Optimization (PSO), for multi-label feature selection tasks. One of the key advantages of ACO is its ability to model the solution space as a graph, allowing it to naturally represent and exploit the similarity and redundancy relationships among features during the search process. This graph-based perspective enables more effective navigation through high-dimensional feature spaces compared to vector-based representations typically used in GA or PSO. Moreover, ACO leverages distributed memory through pheromone trails, which helps maintain information about promising feature subsets across iterations and supports a more balanced exploration–exploitation strategy. These properties have been shown to yield higher-quality feature subsets and more stable convergence behaviour, making ACO particularly well-suited to the challenges of multi-label classification. However, the ACO-based multi-label FS methods aare wrapper methods which required a classifier to evaluate the fature subsets. Most existing ACO-based feature selection methods, e.g. (Hatami et al., 2020b; Tabakhi and Moradi, 2015) utilize a fully connected graph to represent the feature space. However, many real-world applications consist of thousands of features, and searching through such a large, fully connected graph is computationally expensive.

To address the limitations of existing methods, we propose a novel multi-layer graph-based feature selection (FS) framework designed for multi-label classification. Traditional ACO-based FS approaches typically represent the feature space as a fully connected graph, without sufficiently accounting for feature similarity or redundancy. This leads to high computational complexity and suboptimal feature subsets, especially in high-dimensional settings. In contrast, our method constructs a two-layer graph structure in which the first layer groups correlated features using community detection (Louvain algorithm), and the second layer forms a meta-graph representing inter-cluster relationships. A modified ACO algorithm is then applied to this structure, enabling more efficient and redundancy-aware exploration. Additionally, instead of relying on computationally expensive classifier-based evaluations, our framework uses a mutual information-based metric that simultaneously captures relevance to labels and redundancy among features. This design improves both scalability and classification performance across diverse datasets.

- Existing ACO-based FS methods map the solution space using a fully connected graph, with insufficient consideration for the similarity between nodes. In contrast, our method assigns edge weights based on the correlation between corresponding features, which aids in accounting for redundancy during the search process.
- Most existing ACO-based feature selection methods, e.g., (Hatami et al., 2020a; Kashef and Nezamabadi-pour, 2015; Paniri et al., 2019; Tabakhi and Moradi, 2015; Tabakhi et al., 2014), utilize a fully connected graph to represent the feature space. However, many real-world applications consist of thousands of features, and searching through such a large, fully connected graph is computationally expensive. To address this issue, the proposed method constructs a two-layered graph using a community detection method. In one layer, similar features are grouped together, while the other layer contains a meta-graph where each node represents a cluster (Mahmood et al., 2020; Mehrmohammadi et al., 2020, 2021). We then modify the ACO to search through this graph by placing greater emphasis on moving between meta nodes (i.e., communities) and less emphasis on remaining within a community (i.e., a group of similar features). This approach helps to significantly reduce computational complexity (Akhtar et al., 2025).
- A learning model is typically used to evaluate feature subsets in most multi-label feature selection methods, such as those proposed by (Zhang et al., 2017). This can be computationally expensive, especially for real-world applications with large solution spaces. Rather than training a classifier, we propose a novel information-theoretic metric for evaluating feature subsets. Some other multi-label feature selection methods, such as (Hashemi et al., 2020b; Paniri et al., 2019), consider only the relevance of the features to the class labels and neglect the redundancy between features. In contrast, our method evaluates both relevance and redundancy using a mutual information (MI)-based metric. Some multi-label feature selection methods, e.g., (Zhang et al., 2019a), use heuristic searches to find the final solution and lack a strategy to escape from local optima. Our proposed method benefits from the local and global search properties of the ACO search process, which prevents the algorithm from becoming trapped in local optima.
- Our method uses mutual information in the evaluation process, which can capture nonlinear relationships between features and labels without requiring any specific assumptions about the data distribution. This facilitates the identification of characteristics that have a significant influence on the classification model by detecting intricate interactions between features and labels (You et al., 2024; Zhang et al., 2023, 2024b).

The remainder of this paper is organized as follows. Section 2 reviews related work in the area of multi-label feature selection and optimization-based approaches. Section 3 introduces the proposed method, detailing the construction of the feature graph, the multi-layer modeling, and the Ant Colony Optimization-based feature selection process. Section 4 presents the experimental evaluation, including datasets, baseline methods, evaluation metrics, parameter settings, and a comprehensive analysis of the results. Finally, Section 5 concludes the paper and outlines potential directions for future research.

2. Background

2.1. Ant colony optimization (ACO)

ACO is a probabilistic, population-based metaheuristic introduced by Dorigo and Di Caro (1999), inspired by the collective foraging behaviour

of ant colonies in nature. When searching for food, ants deposit a chemical substance called pheromone along the paths they travel. Over time, paths that lead to more better outcomes. Other ants are more likely to follow these pheromone-rich paths, creating a positive feedback loop that reinforces optimal solutions while allowing suboptimal ones to evaporate. This natural mechanism forms the basis of ACO's search and learning process. In its original formulation, ACO was designed to solve combinatorial optimization problems such as the Traveling Salesman Problem (TSP). ACO operates by simulating a colony of artificial ants that construct candidate solutions through a sequence of probabilistic decisions. At each decision point, an ant chooses the next component of its solution (e.g., the next city to visit in TSP) based on two factors: the amount of pheromone present on a path and a heuristic value that estimates the quality of that path. Let $\tau_{ij}(t)$ denote the pheromone level on edge (i,j) at iteration t, and let η_{ij} represent the heuristic desirability of edge (i,j), typically defined as the inverse of the cost or distance. The robability that ant k, currently at node i, selects node j as its next move is given by the following formula:

$$P_{ij}^{k} = \begin{cases} \dfrac{\left[t_{ij}(t)\right]^{\alpha}\left[\eta_{ij}(t)\right]^{\beta}}{\sum_{l\in N_i^k}\left[t_{il}(t)\right]^{\alpha}\left[\eta_{il}(t)\right]^{\beta}} \\ otherwiese \quad 0 \end{cases} \tag{1}$$

where α and β are parameters that control the relative importance of pheromone and heuristic information, respectively, and N_{ik} is the set of feasible nodes that ant k can visit from node i. After all ants construct their solutions, the pheromone trails are updated. First, all pheromone values undergo evaporation to avoid unlimited accumulation and encourage exploration. This is achieved by reducing the pheromone on each edge as $\tau_{ij}(t)\leftarrow(1-\rho)\tau_{ij}(t)$ where $\rho\in(0,1)$ is the pheromone evaporation rate. Then, pheromone is deposited based on the quality of the solutions found by the ants in the current teration. Typically, for each ant k, the pheromone added to the edges it used is given by $\Delta\tau_{ij}^{k}(t)=\frac{Q}{L_k}$ if (i,j) is in ant k's path, otherwise $\Delta\tau_{ij}^{k}(t)=0$. Here Q is a constant and L_k is the cost (e.g., tour length) of the solution found by ant k. The total pheromone update is then computed as $\tau_{ij}(t+1)=\tau_{ij}(t)+\sum_{k=1}^{m}\Delta\tau_{ij}^{k}(t)$ where m is the total number of ants. Through repeated cycles of solution construction and pheromone update, the algorithm onverges toward high-quality solutions as good paths accumulate higher pheromone levels and become increasingly likely to be chosen. The computational complexity of the original ACO algorithm depends on the number of ants N, the number of iterations T, the size of the solution space (e.g., the number of nodes D in a graph), and the cost C of evaluating a single solution. In general, the total time complexity is given by $O(T\times N\times(D+C))$. In classical applications like TSP, the evaluation cost C is linear in the number of cities visited, whereas in other domains such as feature selection, C may involve training classifiers or computing evaluation metrics.

2.2. Community detection: louvain algorithm

Community detection in graph structures is a fundamental task aimed at identifying groups of nodes that are more densely connected internally than with the rest of the network. In feature selection, applying community detection helps group similar features into coherent clusters, thereby reducing redundancy and enabling a more structured exploration of the feature space. The Louvain algorithm (Blondel et al., 2008) is a widely used method for community detection due to its efficiency and scalability. It operates through a modularity optimization process that unfolds in two main phases. Initially, each node in the graph is assigned to its own community. The algorithm then iteratively considers moving each node to the community of one of its neighbors, accepting the move only if it results in a gain in modularity. When no further improvements are possible, the identified communities are collapsed into super-nodes, and the process is repeated on this new meta-graph. This process continues until a stable community structure is reached. The modularity score Q is used to quantify the strength of the division of a network into communities and is defined as:

$$Q=\frac{1}{2m}\sum_{i,j}\left(w_{ij}-\frac{k_i k_j}{2m}\right) \tag{2}$$

where w_{ij} represents the weight of the edge between nodes i and j, k_i and k_j are the weighted degrees of odes i and j, m is the total edge weight in the graph, and $\delta(c_i,c_j)$ is an indicator function that equals 1 if odes i and j belong to the same community and 0 otherwise. The Louvain algorithm is particularly advantageous because of its near-linear complexity with respect to the number of nodes, lack of dependence on manually set parameters such as the number of communities, and its ability to detect hierarchically nested community structures. In our method, this algorithm is used to create a two-layer graph representation in which features are grouped into clusters, and each cluster forms a node in a higher-level meta-graph. This structure enables a more organised and efficient application of the ACO algorithm during feature selection.

3. Related works

Feature selection (FS) is an NP-Hard task, and numerous stochastic optimization methods have been proposed to find near-optimal solutions. Metaheuristics, such as Genetic Algorithm (GA), Artificial Bee Colony (ABC), Particle Swarm Optimization (PSO), and Ant Colony Optimization (ACO), have been widely used to address FS problems. Metaheuristics are stochastic search methods that iteratively find near-optimal solutions. Additionally, some methods inspired by social network optimization have been applied to feature selection tasks (Hamedmoghadam et al., 2018). These are wrapper-based methods that employ a learning algorithm to evaluate the solutions found during iterations. However, these methods often require training a classifier for evaluation, which can be computationally expensive, despite their effectiveness in finding nearly optimal feature sets. This makes them less suitable for real-world problems. To address this issue, information-theoretic measures such as rough set theory and mutual information (MI) have been used to evaluate feature sets (Salem et al., 2022; Sharmin et al., 2019). These methods are particularly advantageous in high-dimensional data scenarios, where they help identify and eliminate redundant or irrelevant features without relying on a specific learning algorithm. Moreover, MI-based approaches can quantify relevance and redundancy, offering a balanced evaluation for feature subsets. However, these techniques also present certain limitations. The computation of mutual information can be expensive, especially when dealing with multi-dimensional interactions. The FS problem is often treated as a single-objective optimization task using the methods mentioned above. However, multiple objectives, such as classifier performance, feature relevance, redundancy, and the number of features, need to be considered when selecting features. When more than one objective is involved, instead of identifying a single optimal solution, a set of non-dominant solutions is obtained. In such cases, measures like crowding distance and niche count are employed to select the final solution. In (Rahmaninia and Moradi, 2018) two novel filter-based online stream feature selection methods based on mutual information named OSFSMI and OSFSMI-k, leveraging mutual information to evaluate feature relevancy and redundancy are introduced. The methods are model-independent, enabling fast processing without the overhead of learning models. Despite their improved performance, potential limitations include sensitivity to noise in streaming data and lack of integration with adaptive learning frameworks. A hybrid parameter adaptation-based ant colony optimization with a dynamic hybrid mechanism named PF3SACO is proposed, combining the PSO for global search (Zhou et al., 2022). The method introduces a dynamic parameter

adaptation mechanism to tune the pheromone importance factor, evaporation rate, and heuristic influence in ACO during the search process. However, the algorithm's complexity and reliance on parameter interactions may pose challenges for scalability and adaptation to domains beyond the traveling salesman problem (TSP).

Several methods in the literature have been proposed for multi-objective feature selection. For example (Zhou et al., 2021), presented a problem-specific non-dominated method based on GA, in which an accuracy-preferred domination operator and three additional operators were introduced to improve convergence speed. The algorithm requires parameter tuning and computational resources typical of evolutionary methods. Recently (Labani et al., 2020), proposed a multi-objective genetic algorithm for text feature selection using the relative discriminative criterion named MORDC. This method avoids dependence on learning models, making it computationally efficient and model-agnostic. Moreover, like other evolutionary algorithms, it may involve a non-negligible computational overhead during the optimization process (Nemati et al., 2024). presents two hybrid feature selection methods based on a GA and the Gray Wolf Optimizer (GWO), enhanced with structured sparse norm-based evaluation. However, these methods may suffer from increased computational costs due to the complexity of the hybrid optimization process. The Semi-ACO method, proposed by (Karimi et al., 2023), employs Ant Colony Optimization for semi-supervised feature selection by maximizing feature relevance and minimizing redundancy among features using a nonlinear heuristic function based on Temporal Difference reinforcement learning. However, it is not suitable for multi-label datasets. A new study proposes a novel multi-objective binary gray wolf optimization for feature selection based on a guided mutation strategy, called MOBGWO-GMS, for feature selection (Li et al., 2023). The method initializes the population based on feature correlation. However, due to its wrapper-based nature and multiple adaptive components, it may impose a higher computational cost (Wang et al., 2025). Introduced a weighted low-rank tensor-based learning (WLTL) framework for unsupervised multi-view feature selection, aiming to address high-dimensional, unlabeled multi-view data challenges. WLTL effectively models complex inter-view relationships through tensor representation and adaptive view weighting, but, the model's performance may depend on the quality of initial spectral clustering.

Table 1, summarizes the properties of the state-of-the-art single-label feature selection methods.

In contrast to single-label FS methods, some applications require classifying each instance into multiple categories. For example, in image annotation tasks, all objects in an image need to be tagged, or in disease classification, the results of an experiment may indicate that a patient has more than one disease. Real-world multi-label classification tasks often involve thousands of features, many of which are irrelevant to the class labels. Including redundant and irrelevant features in the classification process can degrade the classifier's performance. Early multi-label FS methods often converted multi-label datasets into single-label datasets (Spolaôr et al., 2016). However, the major drawback of these approaches is the loss of information during the conversion process. To address this issue, adaptive methods work directly with the entire feature space (Kashef et al., 2018), often in combination with heuristic and metaheuristic techniques. For example (Lee and Kim, 2013), proposed a multi-label FS method based on multivariate mutual information (MI). Additionally (Li et al., 2017; Lin et al., 2016), introduced a neighborhood MI-based metric for multi-label FS, while (Li et al., 2017; Xiong et al., 2021) proposed a granular multi-label FS method that employed the concept of granulation to explore dependencies more effectively. Moreover, utilized a fuzzy MI metric to compute the similarity between features.

Other types of heuristic methods map the problem to a global optimization framework and employ optimization techniques to find optimal solutions. For example (Zhu et al., 2018), utilized an effective norm-2 function to eliminate noisy and irrelevant features while selecting highly discriminative that Integrates feature selection with label recovery in a unified framework, however assumes a linear feature-label relationship, which may limit its performance in highly non-linear domains unless extended. In (He et al., 2019), multi-label learning and feature selection were integrated into a unified framework, addressing the issue of missing labels by incorporating regularization terms into the objective function. Although, the model relies on proper parameter tuning, especially for sparsity control, which may impact stability. Recently (Qian et al., 2022), combined feature and label relevance into a single framework, developing a convex optimization function to guide feature selection based on mutual information.

Table 1
Main properties of existing feature selection methods. **SL**: Single Label, **ML**: Multi Label, **ACO**: Ant Colony Optimization, **GA**: Genetic Algorithm, **ABC**: Artificial Bee Colony, **PSO:** Particle Swarm Optimization**, MI**: Mutual Information, **GWO:** Gray Wolf Optimizer.

Methods	Single Label \ Multi Label	Strategy	Filter/ Wrapper	Multi Objective/ Single Objective
UFSACO (Tabakhi et al., 2014)	SL	ACO	Filter	S O
RRFSACO (Tabakhi and Moradi, 2015)	SL	ACO	Filter	S O
ABACO (Kashef and Nezamabadi-pour, 2015)	SL	ACO	Wrapper	S O
GCACO (Moradi and Rostami, 2015)	SL	ACO	Filter	S O
MBACO (Wan et al., 2016)	SL	GA, ACO	Filter	S O
UPFS (Dadaneh et al., 2016)	SL	ACO	Filter	S O
MGCACO (Ghimatgar et al., 2018)	SL	ACO	Filter	S O
OSFSMI and OSFSMI-k (Rahmaninia and Moradi, 2018)	SL	MI	Filter	S O
ABCoDT (Rao et al., 2019)	SL	ABC	–	S O
TMABC-FS (Zhang et al., 2019c)	SL	ABC	Wrapper	M O
FMABC-FS (Wang et al., 2020)	SL	ABC	–	M O
PS-NSGA (Zhou et al., 2021)	SL	GA	Wrapper, Filter	M O
MORDC (Labani et al., 2020)	SL	GA	Filter	M O
TSHSF-ACO (Ma et al., 2021)	SL	ACO	Hybrid	S O
SemiACO (Karimi et al., 2023)	SL	ACO	Filter	SO
CVSSA (Zhang et al., 2024a)	SL	Competition Mechanism	–	SO
FPO (Zandvakili et al., 2024)	SL	Fuzzy-pathfinder Optimization	–	SO
PF3SACO (Zhou et al., 2022)	SL	PSO, ACO	Hybrid	SO
MOFS-EL (Zhou et al., 2024b)	SL	Ensemble Learning	–	MO
MOBGWO-GMS (Li et al., 2023)	SL	GWO	Wrapper	MO
GWO–ANN–SSN (Nemati et al., 2024)	SL	GA, GWO	Hybrid	MO
AMFEA (Li et al., 2025b)	SL	Evolutionary Algorithm	–	SO
EMSWOA (Miao et al., 2025)	SL	Multi-swarm Whale Optimization Algorithm	–	SO
WLTL (Wang et al., 2025)	SL	1- Weighted Low-rank Tensor Learning	–	SO

This method reduces redundancy and repetitive computations typically associated with heuristic algorithms, however, may be less flexible in capturing nonlinear dependencies compared to deep learning-based feature selection methods. Using matrix factorization, the method proposed by (Hu et al., 2020) extracted common properties between the label and feature matrices, offering a novel approach to feature selection based on shared common modes. Although it provides a unified framework that balances relevance, redundancy, and interpretability, the coupled factorization process may increase computational complexity, particularly for large-scale datasets. In (Zhang et al., 2019a), a manifold regularization approach was introduced for multi-label FS tasks. The problem of multi-label FS has also been studied using stochastic search methods. In (Zhang et al., 2017), the PSO was used to identify prominent features, with the multi-label k-nearest neighbors (MLKNN) classifier employed to evaluate the solutions. As this approach relies on a learning model during the search process, it is a wrapper approach, which incurs a high computational cost. Filter methods have been used to mitigate this issue by employing information-theoretic measures rather than a learning model to evaluate solutions. For instance (Lee et al., 2019), proposed a multi-label FS method for text classification using a memetic algorithm. In (Paniri et al., 2019), a heuristic metric for evaluating feature sets was introduced. Furthermore (Karimi et al., 2023), presented a multi-label FS method for semi-supervised classification that utilizes a nonlinear function and the temporal difference method to learn the heuristic function of Ant Colony Optimization.

All the aforementioned methods are single-objective, focusing on optimizing only one objective in their search processes. However, some recent works have modeled feature selection as a multi-objective optimization task. For instance (Kashef and Nezamabadi-pour, 2019), employed a Pareto-based feature selection method that considered several objectives. In (Paul et al., 2021), a multi-objective swarm intelligence method for multi-label FS was used for streaming data (Asilian Bidgoli et al., 2021). introduced a multi-objective evolutionary FS method that utilized a reference-point strategy to generate diverse solutions. However, these multi-objective feature selection methods often suffer from limitations such as high computational complexity, difficulty in maintaining a balance between relevance and redundancy, and instability in the selected feature subsets across different runs, particularly when applied to high-dimensional or streaming data environments. More recently, the authors (Rafie et al., 2023) proposed an MI-based multi-objective multi-label FS method for streaming features, which assumes that the entire feature space is not available and that features arrive over time. However, the approach may face challenges when deployed in large-scale or time-critical environments, as no specific mechanisms are discussed for real-time acceleration or scalability under high-throughput conditions. In (Zhou et al., 2024b), an ensemble learning technique based on feature relevance-guided selection and multi-objective feature selection (MOFS) was presented. This method builds accurate and diverse individual classifiers using a hybrid MOFS algorithm. However, The use of hybrid MOFS and optimization-based ensemble selection could lead to high computational overhead. Furthermore, a many-objective approach for multi-label feature selection was proposed, taking into account five objective functions that generated numerous non-dominated solutions.

Multi-label feature selection considering join mutual information and interaction weight (MFSJMI) method (Zhang et al., 2023), enhances multi-label feature selection by incorporating joint mutual information and interaction weights to account for label correlations. It improves upon traditional assumptions by offering a more realistic label distribution analysis, although, the use of high-order mutual information and label correlation modeling may increase computational costs. The method in (Li et al., 2022) leverages a self-expression model and L2, 1-norm to handle label correlations and remove redundant information. However, The method may struggle with scalability on extremely large-scale datasets due to the reliance on matrix operations involving label and feature spaces. Stable label relevance and label-specific features (Yang et al., 2023) called (LRLSF), proposed a multi-label feature selection method that Combines both global and local label relevance, offering a more complete understanding of label relationships. However, The combination of multiple regularization terms and representation strategies may increase training time (Faraji et al., 2024). proposed a multi-label feature selection method that identifies discriminative features by utilizing both global and local label correlations. Explicit modeling of both global and local label correlations improves feature relevance across labels, however, it has potential sensitivity to hyperparameters in the regularization terms and optimization algorithm. A mutual information-based multi-label feature selection method is proposed in (Zhang et al., 2024b), which employs a label augmentation algorithm to convert logical label vectors into distributions that better reflect correlations. This method effectively captures mutual information between features and label distributions, however, the algorithm relies on the quality of neighborhood information, which may degrade if feature space relationships are noisy or poorly defined. A new feature selection method (Zhang et al., 2024a) combines a modified salp swarm algorithm with variable shifted windows and a competitive mechanism to enhance feature selection. this strategy enhances local search capability, but, the approach focuses on batch-mode datasets and may require adaptation for streaming or large-scale real-time data. Multi-label multi-view learning is a domain in which data are dispersed over numerous perspectives, each providing different semantic representations (Hao et al., 2024). Unlike prior work, this method captures the diversity of label distributions across different views, addressing the label space inconsistency problem, however, Its performance on highly noisy or extremely imbalanced multi-view datasets remains to be thoroughly assessed. A new feature selection method for label distribution learning (LDL) data (You et al., 2024) combines fuzzy mutual information with statistical data distribution. The method effectively integrates fuzzy logic into mutual information calculations, allowing better handling of uncertainty inherent in LDL data, but, Further exploration may be needed to assess performance on LDL tasks with highly imbalanced or sparse label distributions. A feature selection method developed by (Ding et al., 2024) employs neighborhood mutual information and label disambiguation. This method efficiently addresses continuous features while avoiding issues related to data discretization. Although, the performance of the method may be sensitive to the neighborhood size or granularity, requiring careful tuning. The feature selection method introduced by (Zandvakili et al., 2024) utilizes fuzzy-pathfinder optimization (FPO) to enhance feature relevance. The use of correlation-based criteria for initial population generation increases the likelihood of better convergence paths, but, Like other population-based methods, FPO does not provide theoretical convergence guarantees to the global optimum. Recently (Dai and Wang, 2025) proposed a granularity measure for capturing feature interactions and a symmetric discriminant weight for evaluating feature-label correlations in multi-label feature selection. The proposed measures ensure theoretically desirable properties, such as monotonicity, improving the robustness of the selection process. However, as with many feature selection algorithms, performance may depend on carefully tuned thresholds or granularity levels, which might limit ease of use. A strongly relevant label gain and label mutual aid called SRLG-LMA (Dai et al., 2024) proposes a novel multi-label feature selection method that explicitly incorporates both strongly relevant label gain and label mutual aid that breaks away from the common assumption of uniform label relevance. Although, the integration of multiple label-aware components, may introduce additional computational overhead. Multi-label feature selection based on minimizing feature redundancy called MFS-MFR (Zhou et al., 2024a) introduces a novel approach for multi-label feature selection by leveraging mutual information estimation to capture nonlinear correlations between features and labels. A key strength of the approach lies in its ability to model both linear and nonlinear relationships, providing a more expressive representation of feature-label

interactions. LEFMIFS (Label Enhancement and Fuzzy Mutual Information-based Feature Selection) (Yin et al., 2024) is a robust multi-label feature selection algorithm proposed to address challenges associated with high-dimensional and noisy data that enhances robustness and interpretability in noisy, real-world multi-label datasets, however, the method introduces additional computational complexity due to the dual-space and fuzzy rough set integration, which may limit scalability for very large datasets. A Multi-label feature selection with feature reconstruction and label correlations named (FSFL) (Zhang et al., 2025) introduces a novel multi-label feature selection method that integrates feature reconstruction and label correlations to address the challenges posed by high-dimensional data common in multi-label tasks by constructing a low-dimensional embedding that preserves the manifold structure of the original feature space. This method Effectively captures label correlations often overlooked in traditional methods, however, Effectiveness may vary depending on the intrinsic structure of the data manifold. Recently, an enhanced multi-label feature selection considering label-specific relevant information named LSRIFS was proposed (Han et al., 2025) unlike previous methods, it assesses relevance from both macro and micro perspectives, ensuring that features selected are not only collectively relevant but also strongly associated with individual labels. This method addresses the limitations of greedy algorithms by incorporating both global and individual label correlations, although, the performance may depend on the nature of label distribution and dataset sparsity.

Table 2, summarizes the properties of the state-of-the-art multi-label feature selection methods.

4. Proposed method

This section aims to provide a detailed description of our multi-label FS method. The proposed method employs various notations and parameters to model the multi-label feature selection process. Table 3 presents a summary of the key symbols and their descriptions to enhance clarity and facilitate understanding of the subsequent analysis.

Let $X \in R^n$ denotes the instance space where each instance x_i is typically described by assigning values to n features $F = \{f_1, f_2, \ldots, f_n\}$, and $L = \{l_1, l_2, \ldots, l_q\}$ represents the label space includes q distinct labels. Given a multi-label dataset $D = \{x_i, Y_i | 1 \leq i \leq m\}$ containing m samples where $Y_i \in \{0,1\}^{|L|}$ defines the set of labels that i^{th} the sample belongs. The goal of multi-label FS is to decrease the size of the feature space by identifying a subset of the most relevant $k \ll n$ features. The selected feature set should be discriminative enough to predict the class labels accurately. It is crucial to determine the features that have a strong association with the class labels. Moreover, redundant features—those correlated with other selected features—do not improve the discriminatory ability of feature sets and increase the model complexity. The proposed method aims to identify a set of non-similar features that are highly correlated with the target class. In the case of choosing k features, the search space contains the number of $\binom{n}{k}$ feature sets with many local minimal. Our approach is based on the ACO algorithm, which has been successfully applied to various optimization tasks. The proposed method aims to improve its efficiency in finding and selecting relevant and non-redundant features in multi-label classification tasks. MR2MLFS is divided into four parts shown in Fig. 1: (1) A structured pipeline beginning with the construction of a feature similarity graph derived from correlation analysis, (2) A Louvain community detection algorithm is then applied to create a multi-layered graph, (3) The ACO is employed to explore the graph, balancing local and global search by moving within and between clusters, and (4) Feature subsets are evaluated using a mutual information-based fitness function, and features are ranked based on pheromone levels. The top-ranked features are selected and validated using the ML-kNN classifier, with performance assessed through standard multi-label metrics. All the steps of the

Table 2
Main properties of existing feature selection methods. **SL**: Single Label, **ML**: Multi Label, **ACO**: Ant Colony Optimization, **GA**: Genetic Algorithm, **ABC**: Artificial Bee Colony, **PSO:** Particle Swarm Optimization, **MI**: Mutual Information.

Methods	Single Label \ Multi Label	Strategy	Filter/ Wrapper	Multi Objective/ Single Objective
MFNMI (Lin et al., 2016)	ML	Neighborhood MI	Filter	S O
MPSOFS (Zhang et al., 2017)	ML	PSO	–	M O
SCLS (Lee and Kim, 2017)	ML	MI	Filter	S O
MLMLFS (Zhu et al., 2018)	ML	Iterative Reweighted Least Squares	Filter	S O
MCLS (Huang et al., 2018)	ML	Manifold Learning	Filter	S O
CSCC (Teisseyre et al., 2019)	ML	Penalized Logistic Regression	Wrapper	SO
PFS (Kashef and Nezamabadi-pour, 2019)	ML	Pareto Dominance Concept	Filter	M O
MICO (Sun et al., 2019)	ML	MI, Convex Optimization	Filter	S O
MLLF (He et al., 2019)	ML	Label Correlation	–	S O
LRFS (Zhang et al., 2019b)	ML	Label Redundancy, Conditional MI	Filter	S O
MDFS (Zhang et al., 2019a)	ML	Manifold Regularization	embedded	S O
MLACO (Paniri et al., 2019)	ML	ACO	Filter	S O
CLSF (Che et al., 2020)	ML	Label Correlation	–	S O
MGFS (Hashemi et al., 2020b)	ML	Correlation Distance, Page Rank	Filter	S O
LDMI (Qian et al., 2020)	ML	Label Distribution, MI	Filter	S O
SCMFS (Hu et al., 2020)	ML	Matrix Factorization	Embedded	S O
SUMLFS (Dai et al., 2020)	ML	Fuzzy MI, Symmetric Uncertainty	Filter	SO
MFS-MCDM (Hashemi et al., 2020a)	ML	Multi-Criteria Decision Making	Filter	S O
MMFS (Dong et al., 2020)	ML	NSGA III	Wrapper	Many O
MMOFS (Paul et al., 2021)	ML	PSO, Online FS	Filter	M O
Ant-TD (Paniri et al., 2021)	ML	Reinforcement Learning, ACO	Filter	S O
FSLE (Xiong et al., 2021)	ML	Label Distribution, Fuzzy MI	Filter	S O
SFAM (Lv et al., 2021)	ML	Manifold Regularization	Embedded	S O
LFFS (Fan et al., 2022)	ML	Label Correlations	Embedded	S O
MFSJMI (Zhang et al., 2023)	ML	Join MI	Filter	SO
LRLSF (Yang et al., 2023)	ML	Stable Label Relevance	Filter	SO
MLFS-GLOCAL (Faraji et al., 2024)	ML	Global and Local Label Correlation	Filter	SO
SFSMIE (Zhang et al., 2024b)	ML	Label Enhancement Technique	–	SO

(continued on next page)

Table 2 (*continued*)

Methods	Single Label \ Multi Label	Strategy	Filter/ Wrapper	Multi Objective/ Single Objective
I2VSLC (Hao et al., 2024)	ML	View-specific Label Relationships	–	SO
LDFM (You et al., 2024)	ML	Label Distribution and Fuzzy MI	–	SO
PLFSDN (Ding et al., 2024)	ML	Neighborhood MI	–	SO
SRLG-LMA (Dai et al., 2024)	ML	label gain and label mutual aid	Filter	SO
MFS-MFR (Zhou et al., 2024a)	ML	MI	Embedded	SO
LEFMIFS (Yin et al., 2024)	ML	Label Enhancement, Fuzzy MI	Filter	SO
MLDMF (Dai and Wang, 2025)	ML	Granularity Information	–	SO
Q&A (Li et al., 2025a)	ML	Multi-scale Feature Extraction	–	SO
FSFL (Zhang et al., 2025)	ML	Label Correlations	–	SO
LSRIFS (Han et al., 2025)	ML	Label-Specific	–	SO

Table 3
Description of symbols and notations used in the proposed method.

Symbol	Description
X	Multi-label dataset with p instances and n features
NC	Number of ACO cycles (iterations)
NA	Number of ants used in the ACO process
m	Size of the final selected feature subset
n	Total number of original features in the dataset
ε	Initial pheromone value assigned to edges in the graph
ρ	Pheromone evaporation rate
r	Parameter for deciding whether to stay in the current feature cluster
θ	Threshold determining the minimum weight for connecting edges in the graph
MI(f_i, l_j)	Mutual Information between feature f_i and label l_j
CL(f_i, l_j)	Correlation level between feature f_i and label l_j
EM(f_i, f_j)	Euclidean distance between correlation vectors of features f_i and f_j
Q(S_k)	Evaluation function representing the overall quality of the feature subset selected by the *k*-th ant
$RR_k(f_i)$	Relevance-redundancy score of feature f_i in the subset selected by ant *k*
δ(k, f_i)	An indicator function: 1 if feature f_i is selected by ant *k*, 0 otherwise
Δτ(f_i)	Total pheromone value assigned to feature f_i, based on its contribution across all ants
SoF	Sorted list of features based on relevance and redundancy criteria
SF	Final selected feature subset after optimization

proposed method presented in this paper, are summarized and integrated in Fig. 1.

The existing swarm intelligence-based feature selection (FS) methods primarily focus on feature relevancy during their search processes. The proposed method aims to extend this by incorporating both relevancy and redundancy of features in the search process of Ant Colony Optimization (ACO). We modify the ACO's exploration and exploitation strategies to assign higher-ranked values to non-similar features that are highly correlated with the target class. In this method, features are represented as nodes in a weighted, undirected graph, where the strength of the connection between two features is denoted by the edge weight. A two-layered graph is then generated: one layer groups similar features through graph clustering, creating a meta-node for each group. The second layer represents the relationships between these meta-nodes (clusters of features). The ACO is then applied to this two-layered graph. In the first step, ants select a meta-node, and they must decide whether to remain within that node or move to another. If they remain, they proceed to the second layer to explore individual features within the meta-node. To ensure non-redundant features are selected, the ACO is modified so that ants remain within a cluster with a low probability and traverse between meta-nodes with a higher probability, promoting diversity. The method also aims to identify features that are strongly correlated with the target labels features that contain sufficient information to predict the target classes. To achieve this, the pheromone evaporation mechanism is adjusted so that relevant features are assigned higher pheromone values. Additionally, we introduce a novel mutual information (MI)-based metric to evaluate feature subsets. This metric assesses the correlation between the feature set and the class labels, while also accounting for internal similarities between features.

Fig. 2 provides a step-by-step visual representation of the ptoposed method. The workflow of the proposed method begins with a multi-label dataset in which each instance is associated with multiple class labels. The objective is to identify a subset of features that are highly relevant to these labels while minimizing redundancy. First, a similarity matrix is computed between all pairs of features using an information-theoretic criterion, such as mutual information, to quantify how strongly features are related to one another. This matrix forms the basis for constructing a fully connected graph, where each node represents a feature and each edge weight corresponds to the computed similarity value, capturing the structural relationships between features. Next, a graph clustering algorithm, specifically Louvain community detection, is applied to group similar features into communities. These clusters represent sets of statistically correlated features that serve as high-level representations of the feature space. A multi-layer graph is then generated, consisting of a feature layer containing individual nodes connected by similarity and a cluster layer representing inter-cluster relationships. The ACO algorithm is adapted to operate over this multi-layered graph, with ants probabilistically exploring the graph by moving within clusters with lower probability (exploitation) and moving between clusters with higher probability (exploration). This strategy ensures a balance between local refinement and global search, leading to the selection of highly relevant and minimally redundant features. Finally, features are ranked according to the accumulated pheromone values, and the top-ranked features are selected as the final output. This workflow not only clarifies the operation of the proposed method but also highlights its novel components, including the dual-layer graph representation and the customised ACO-based search mechanism.

4.1. Creating a graph of the features

This step represents the feature space as a fully connected graph $G = < V, E, W >$. Each node $u_i \in V$ represents a feature $f_i \in F$ and each edge $e_{ij} \in E$ shows an edge between f_i and f_j. Fig. 3 illustrates an example of multi-label data where each instance is associated with multiple class labels. It visually demonstrates how features and labels are organized to support multi-label feature selection.

Inspired by (Hashemi et al., 2020b), the similarities between features and labels are considered when defining the edge weights. Two features are considered correlated if they share similar information for predicting class labels. To quantify this correlation, we employ mutual information, a widely used measure, calculated as follows:

$$MI(f_i, l_j) = \sum_{a \in f_i} \sum_{b \in l_j} p(a,b) log \frac{p(a,b)}{p(a)p(b)} \tag{3}$$

where $p(a)$ and $p(b)$ are marginal probability density functions, and $p(a,b)$ denotes the joint probability between a and b, mutual information measures non-linear relationships between two features. It quantifies how much information about one random variable can be gained by observing another. In other words, if two features f_i and l_j are dependent, knowing something about f_i provides additional information about l_j. A

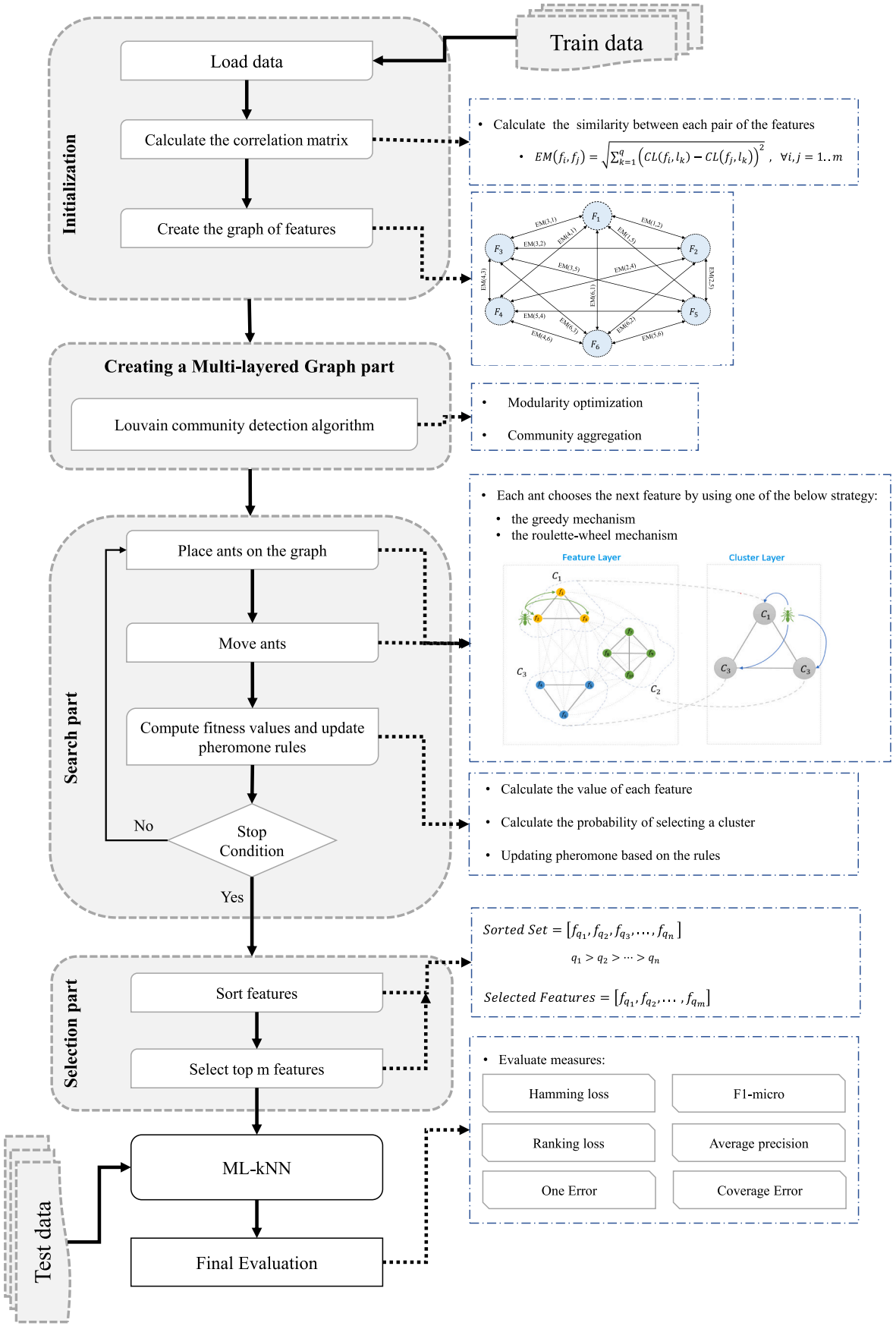


Fig. 1. The flow graph of the main steps of the proposed method.

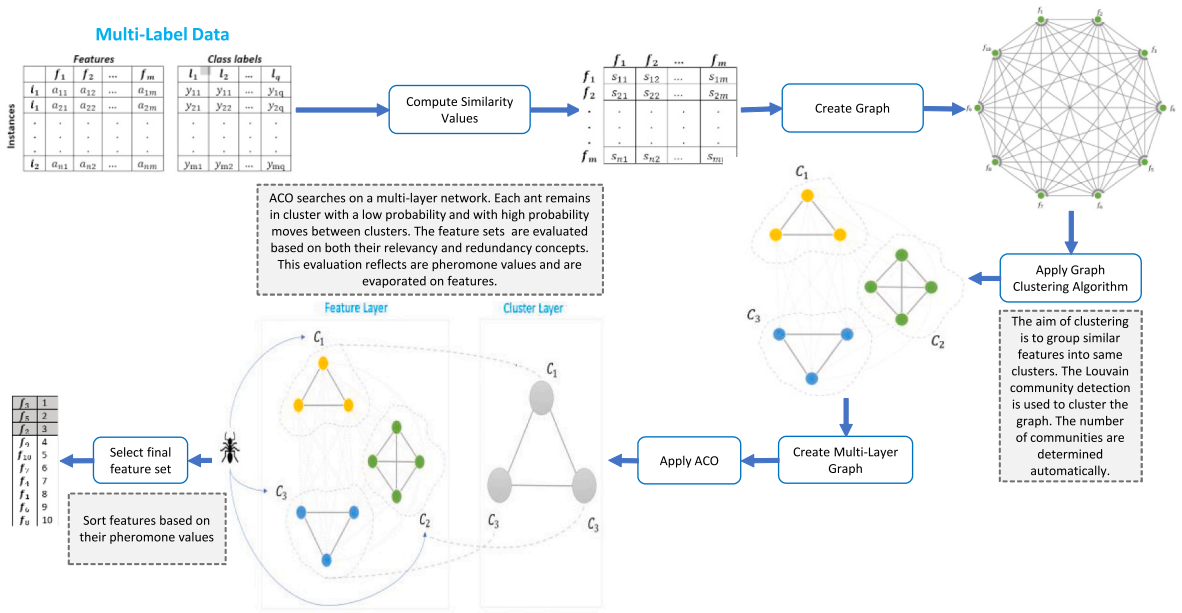


Fig. 2. The illustrative example of the proposed methodology Steps: The procedure begins by computing the similarity between multi-label data instances. Subsequently, a feature graph is constructed based on these similarities. In the next step, the Louvain community detection algorithm is applied to this complete graph. Finally, the Ant Colony Optimization algorithm is employed to select the most relevant features.

Features

Instances	f_1	f_2	...	f_m
i_1	a_{11}	a_{12}	...	a_{1m}
i_1	a_{21}	a_{22}	...	a_{2m}
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
i_n	a_{n1}	a_{n2}	...	a_{nm}

Class labels

l_1	l_2	...	l_q
y_{11}	y_{11}	...	y_{1q}
y_{21}	y_{22}	...	y_{2q}
.	.	.	.
.	.	.	.
.	.	.	.
y_{n1}	y_{n2}	...	y_{nq}

Fig. 3. Multi-label data.

high mutual information value indicates a high correlation between two variables, whereas a low value indicates a weak correlation. If mutual information is zero, two random variables are independent. The similarity between features and labels is computed as follows:

$$CL(f_i, l_j) = MI(f_i, l_j)\ \forall i = 1..m, j = 1..q \tag{4}$$

where $f_i \in F$ shows a feature and $l_j \in L$ denotes a label, each feature is characterized by a vector of its correlation with all labels. The correlation between two features f_i and f_j is defined as the Euclidean distance between their respective vectors, as follows:

$$EM(f_i, f_j) = \sqrt{\sum_{k=1}^{q} \left(CL(f_i, l_k) - CL(f_j, l_k)\right)^2}, \forall i, j = 1..m \tag{5}$$

This correlation is subsequently used to define the edge weights in the graph. A step-by-step example is provided to demonstrate how the graph is generated from the multi-label data. Consider a multi-label dataset containing four instances, six features, and three labels, as follows:

$$X = \begin{bmatrix} 0 & 47 & 1 & 4 & 1 & 5 & 1 & 1 & 0 \\ 3 & 1099 & 6 & 0 & 3 & 2 & 0 & 1 & 1 \\ 5 & 600 & 1 & 0 & 2 & 4 & 0 & 1 & 0 \\ 0 & 8512 & 119 & 1 & 3 & 2 & 0 & 0 & 1 \end{bmatrix}$$

The correlation matrix (computed using Eq. (4)) is as follows:

$$CL = \begin{bmatrix} 0.2821 & 0.2821 & 0.4082 \\ 0.6368 & 0.6368 & 0.7071 \\ 0.2821 & 0.7354 & 0.8164 \\ 0.7354 & 0.7354 & 0.4082 \\ 0.7354 & 0.2821 & 0.4082 \\ 0.7354 & 0.2821 & 0.8164 \end{bmatrix}$$

In this representation, the rows correspond to features, and the columns represent labels. For example, to obtain the value of $CL(3,2)$, the mutual information (MI) between f_3 and l_2 is calculated as follows:

$$CL = MI(f_3, l_2) = \begin{bmatrix} 1 \\ 6 \\ 1 \\ 119 \end{bmatrix} [1\ 1\ 1\ 0] = 0.7354$$

Then, the correlation matrix is obtained using Eq. (5) as:

$$EM = \begin{bmatrix} 0 & 0.5839 & 0.61 & 0.6410 & 0.4533 & 0.61 \\ 0.5839 & 0 & 0.3840 & 0.3298 & 0.4742 & 0.3840 \\ 0.61 & 0.3840 & 0 & 0.61 & 0.7599 & 0.6410 \\ 0.6410 & 0.3298 & 0.61 & 0 & 0.4533 & 0.61 \\ 0.4533 & 0.4742 & 0.7599 & 0.4533 & 0 & 0.4082 \\ 0.61 & 0.3840 & 0.6410 & 0.61 & 0.4082 & 0 \end{bmatrix}$$

E.g. the value of $EM(f_2, f_3)$, is equal to the Euclidean distance between f_2 and f_3 and the details are provided as follows:

$$EM(f_2, f_3) = \sqrt{(0.6368 - 0.2821)^2 + (0.6368 - 0.7354)^2 + (0.7071 - 0.8164)^2} = 0.3840$$

Finally, to generate the graph, each feature is considered as a node, and the weights between edges are the correlation between corresponding features (i.e. $w_{ij} = EM(f_i, f_j) \forall i,j = 1..m$). Fig. 4 illustrates the constructed similarity graph, where each node represents a feature and directed edges denote computed similarity (EM) values between feature pairs. This graph structure forms the basis for subsequent clustering and feature selection steps in the proposed method.

4.2. Creating a multi-layered graph

The objective of this step is to construct a multi-layered graph. In the first layer, correlated features are grouped, and a meta-graph is then created on top of this layer. Each node in the meta-graph corresponds to a cluster from the first layer. Thus, the first layer consists of several separate fully connected graphs, each containing similar features. This structure enables the algorithm to account for feature redundancy more effectively. To group similar features, we employ a community detection method. Specifically, we utilize the well-known Louvain community detection algorithm to divide the graph into communities of highly correlated features (Blondel et al., 2008).

The Louvain community detection method has two main steps: modularity optimization and community aggregation. The modularity is defined as follows:

$$Q = \frac{1}{2m} \sum_{ij} \left(w_{ij} - \frac{k_i k_j}{2m} \right) \delta(c_i, c_j) \tag{6}$$

where w is the weight matrix and w_{ij} represents the weight between nodes i and j, $k_i = \sum_j w_{ij}$, c_i is a community that nide i belongs to, and

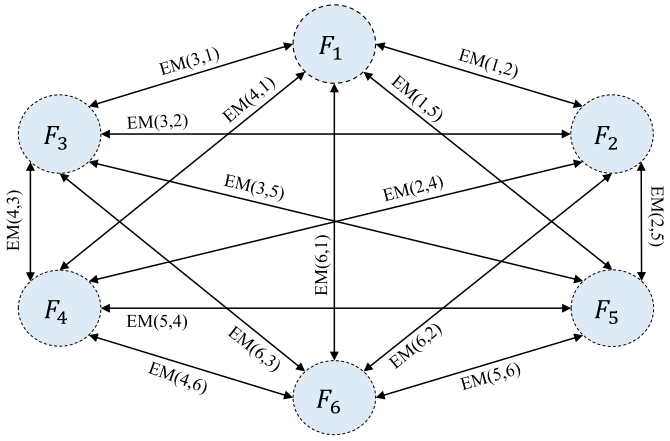


Fig. 4. The generated graph from the multi-label data.

$\delta(c_i, c_j)$ is equal to zero if both i and j belong to the same community, otherwise it is set to 1. $m = \sum_{ij} w_{ij}$ is the sum of the weights of all edges in the graph. Using this definition, the modularity of each community c_j is defined as follows:

$$Q_j = \frac{\sum_{k,l \in c_j} w_{kl}}{2m} - \left(\frac{\sum_{k \in c_j, f \in V} w_{kf}}{2m} \right)^2 \tag{7}$$

The first term represents the sum of interconnection relations for c_j, while the second term captures the sum of its outgoing connections. To maximize modularity, at the initial level, each node is treated as its own community. Then, for each node i, it is first removed from its current community and then considered for addition to neighboring communities. A modularity-based method is applied to the neighbors of each node, determining whether the node should be moved from its current community to another. If no improvement in modularity is observed, the node remains in its original community; otherwise, it is merged with the neighboring community that yields the highest modularity improvement. These steps are repeated iteratively in cycles until no further significant improvements in modularity are detected. Additional details can be found in (Blondel et al., 2008). Fully connected graphs often contain noise and binarizing them is a good way to denoise. To binarize a fully connected graph, edges with weights lower than a predefined threshold $\theta \in [0\ \ 1]$ are removed. The community detection algorithm is then applied to the resulting binary graphs. The optimal value of θ is obtained through sensitivity analysis, in which several values of this parameter are evaluated, and the one yielding the best results is selected.

4.3. ACO-based search

We extend the classical ACO to make it applicable to a multi-layered graph. ACO searches the graph for highly relevant, non-redundant features and allocates high pheromone values to them. Before the algorithm begins, the initial pheromone values of the features and the size of the final feature set are determined. Each ant is then randomly positioned on a node of the graph to initiate its journey. To this end, a cluster is first chosen randomly, and the ant is subsequently placed randomly on one of the nodes within that cluster. A transition rule is proposed to assist the ants in traversing the multi-layered graph, whereby similar (dis-similar) features are selected with a low (high) probability. Fig. 5 illustrates how an ant moves within a cluster or traverses between different clusters. The ant must decide whether to remain in its current cluster (C1, C2, or C3) or switch to a different one. If the ant opts to stay in the current cluster (the feature layer part), it must select the next feature from the cluster using a greedy or Roulette-wheel method.

State transition rule: To choose one of the search mechanisms, a random number $q \in [0\ 1]$ is generated. If this number is lower than a predefined threshold q_0, the greedy mechanism is employed, otherwise, we use the roulette-wheel mechanism to choose the next feature. Eq. (8) is used when the k_{th} ant chooses the next feature f_j:

$$f_j = argmax_{f_i \in UF^{C_i}} \left\{ [\tau_{f_i}] [\eta(f_i, VF_k)]^{\beta} \right\}, if q \le q_0 \tag{8}$$

where UF^{C_i} represents the unvisited features from the current cluster C_i, τ_{f_i} denotes the pheromone value of the feature f_i, VF_k is denoted the visited features by ant k, and $\eta(F_v, VF_k)$ indicates the heuristic desirability information. $\beta > 0$ is a parameter that controls the importance of

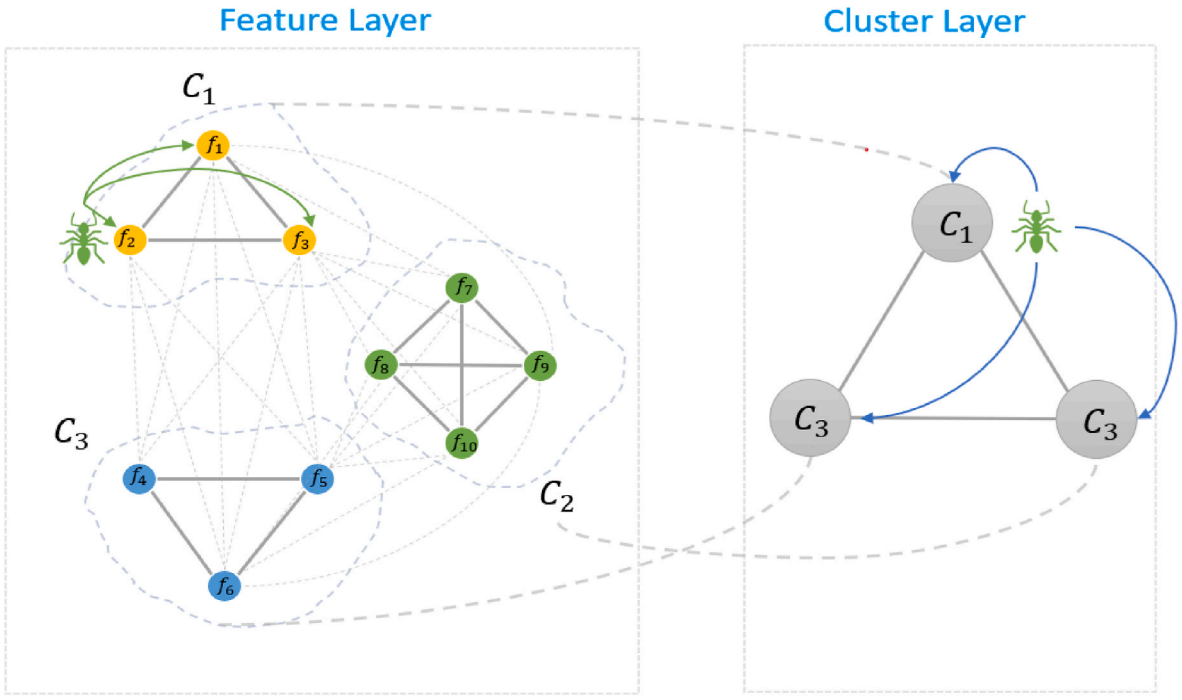


Fig. 5. Two-layer graph structure for feature selection.

the desirability information. In the case of choosing the Roulette-Wheel selection mechanism, the following equation (i.e. Eq. (9)) is used to choose the next feature:

$$P\left(f_j, VF_k\right) = \frac{\left[\tau_{f_j}\right]\left[\eta\left(f_j, VF_k\right)\right]^{\beta}}{\sum_{f_i \in UF^{C_i}} \left[\tau_{f_i}\right]\left[\eta(f_i, VF_k)\right]^{\beta}}, ifq > q_0 \tag{9}$$

To implement the roulette-wheel mechanism, the probabilities of all potential features are first calculated, after which the roulette-wheel selection is applied to these probabilities to determine the next feature. These equations provide a well-balanced combination of exploration and exploitation strategies. It is crucial to emphasize that the exploration component of the search process prevents the ants from converging on a single path.

Ants must select features that are both non-redundant and highly relevant. To achieve this, a specific desirability rule is introduced in this paper. This rule encourages the ants to choose features that exhibit minimal redundancy with previously selected features while demonstrating the highest dependence on the set of labels. This rule is articulated in the following equation (Eq. (10)):

$$\eta(f_i, VF_k) = \sum_{l \in L} MI(f_i, l) - \frac{1}{|VF_k|} \sum_{f_m \in VF_k} EM(f_i, f_m) \tag{10}$$

where $EM(f_i, f_u)$ is the weight of the edge between f_i and f_m that also shows the redundancy between features.

If an ant opts to leave its current cluster, it randomly selects a feature from an unvisited cluster. When transitioning to a different cluster, the selection of the next cluster is influenced by its potential to yield non-redundant features that are highly dependent on the class labels. The pheromone levels associated with the features encapsulate these concepts, and the sum of all pheromone values within a cluster serves as a critical measure for selecting the next cluster. The probability of selecting a cluster C_q is defined as follows:

$$P(C_q, VC_k) = \frac{\left[\tau_{C_q}\right]\left[\gamma(C_q, VC_k)\right]^{\beta_c}}{\sum_{C_m \in UC} \left[\tau_{C_m}\right]\left[\gamma(C_m, VC_k)\right]^{\beta_c}} \tag{11}$$

where VC_k is a set of clusters visited by ant k, β_c is a parameter that controls the importance of prior knowledge in choosing the next cluster. τ_{C_q} is defined as the summation of all pheromone values of the features within cluster C_q (i.e. $\tau_{C_q} = \sum_{f_i \in C_q} \tau_{f_i}$) and $\gamma\left(C_q, VC_k\right)$ shows the desirability of choosing a cluster and is defined follows:

$$\gamma\left(C_q, VC_k\right) = \frac{1}{|C_q|} \sum_{f_i \in C_q} \sum_{l \in L} MI(f_i, l) - \frac{1}{|C_q||VF_k|} \sum_{f_i \in C_q} \sum_{f_m \in VF_k} EM(f_i, f_m) \tag{12}$$

The first term shows the relevance of the cluster to the target labels, while the second term represents the average redundancy between the selected features and those in the cluster C_q. In other words, $\gamma\left(C_q, VC_k\right)$ shows the preference of choosing the cluster C_q as prior knowledge when pheromone values are insufficient to guide the ants effectively.

Pheromone updating rules: When all ants finish their tours, the quality of their paths is evaluated and reflected as the pheromone values. In this case, we utilize an updating rule to revise the pheromone values of the features. This process is repeated for a predetermined number of iterations. Finally, the top-k features are chosen as the final set. The following equation is used to evaluate the quality of a feature $f_i \in S_k$ founded by ant k as follows:

$$RR_k(f_i) = \sum_{l_m \in L} \sum_{\substack{f_j \in S_k \\ f_i \neq f_j}} \alpha MI(f_i, l_m) - \gamma MI\left(f_i, f_j\right) \tag{13}$$

where L is a set of labels, and γ as a factor that regulates the significance of pertinence in comparison to duplicity. The evaluation of the feature set discovered by each ant is calculated as follows:

$$Q(S_k) = \sum_{f_i \in S_k} RR_k(f_i) \tag{14}$$

The final quality of a feature is determined by summing the quality values of all the sets of features to which it belongs. The overall quality of the feature is represented as a pheromone value for f_i:

$$\Delta\tau(f_i) = \sum_{k=1}^{|ants|} \delta(k, f_i) Q(S_k) \tag{15}$$

where $\delta(k, f_i) = 1$ if f_i is included in the set of features identified by the k-th ant. Finally, the pheromone of a feature f_i is then updated as follows:

$$\tau^{t+1}(f_i) = (1 - \rho)\tau^t(f_i) + \Delta\tau^{t+1}(f_i) \tag{16}$$

The pheromone value of f_i at time t is denoted by $\tau^t(f_i)$, while the evaporation coefficient is represented by ρ. Using the proposed search strategy, a high pheromone value indicates features that are highly dependent on a set of labels as well as a non-redundant set of features.

4.4. Feature ranking and selection

Finally, to choose top-rank features for classification, the features must first be sorted in descending order of pheromone values, then the proposed method selects the features with a higher amount of pheromone. For example, suppose $q_1 > q_2 > \ldots > q_n$ are pheromones of the features and sorted based on their values in descending order; in that case, sorting features (SoF) will be as follows:

$$\begin{aligned} &SoF = \left[f_{q_1}, f_{q_2}, f_{q_3}, \ldots, f_{q_n}\right] \\ &q_1 > q_2 > \ldots > q_n \end{aligned} \tag{17}$$

In fact, the ants seek to give more pheromone to features with greater $\Delta\tau$ values; according to this, the important features with the highest relevancy to the labels set and lowest redundancy to the other features have more chance to choose at the end. Consider the above example; the final features subset (SF), given that the number of features is m, will be as follows:

$$SF = \left[f_{q_1}, f_{q_2}, f_{q_3}, \ldots, f_{q_m}\right] \tag{18}$$

The features are ranked in descending order based on their pheromone values, with those at the top of the ranking list selected as the solution. Further details regarding these steps are provided in Algorithm 1.

4.5. Computational complexity analysis

Our method consists of four steps. The initial step involves calculating the correlation between the features and the labels, which has a computational complexity of $O(pn|L|)$, where |L|, p, and n are referred to the number of labels, samples, and features, respectively. Next, we calculate the correlation between each pair of features using Euclidean distance, which requires $O\left(pn^2|L|\right)$. Overall, the complexity of the first step is $O\left(pn|L| + n^2p|L|\right)$, which simplifies to $O\left(n^2p|L|\right)$. The community detection algorithm used in the second step has a complexity of $O(nlog\, n)$. In the third step, the ACO search strategy runs iteratively for NC times. Each ant selects the next feature using the Roulette-Wheel strategy to choose m features, with each decision requiring $O(np)$ time for computing mutual information. Additionally, it is necessary to compute pheromone values for each ant using Eqs. (13)–(16) which requires $O(p|L|)$. If the ants are run in parallel, the complexity of this step is $O(NCnp)$. Finally, the fourth step involves sorting the features, which

requires $O(nlogn)$ time. Thus, the overall complexity of the method is $O(n^2p|L| + nlogn + NCnp + nlogn)$, which reduces to $O(n^2p|L| + NCnp)$.

Algorithm 1. Maximum Relevant Minimum Redundant Multi-Label Feature Selection using Ant Colony Optimization

Input $\boldsymbol{X}$: A multi-label dataset with n features, p instances, and |L| labels, and |F| features, $\boldsymbol{NC}$: #cycles, $\boldsymbol{NA}$: # ants, $\boldsymbol{m}$: the size of final features, $\boldsymbol{n}$: the size of original features, $\boldsymbol{\epsilon}$: Initial value of pheromone, $\boldsymbol{\rho}$: evaporation rate of pheromone, $\boldsymbol{r}$: A parameter for deciding to remain in the current cluster, $\boldsymbol{\theta}$: A parameter that determines the minimum size of the edges

Output Feature subset $S = \{f_1, f_2, \dots, f_m\}$

Begin algorithm
  Compute the correlation between features and labels (CL) using Eq. (4)
  Compute feature similarity matrix using Eq. (5)
  Create a graph of features
  Remove edge weights that are less than θ
  Clustering the graph includes n original features and k clusters
  $\tau[i] = C, \quad i = 1 \dots n$
  For t = 1 to NC **do**
    $Ant_{score}[i] = 0 i = 1 \dots NA$
    $\Delta\tau[j] = 0 j = 1 \dots n$
    For w = 1 to NA **do**
      $Visited_{cluster} = \emptyset$, $unvisited_{cluster} = \{cluster\ 1, \dots, cluster\ k\}$
      While ($|Visited_{cluster}| < k$) **do**
        Place the ant w randomly in one of the clusters in the $unvisited_{cluster}$
        Select the next feature f in $current_{cluster}$
        Move the k_{th} ant to the new selected feature f
        Generate a random value (Rand is between [0,1])
        If ($Rand > r$) **then**
          do while ($Rand > r$)
            Select one of the features in $current_{cluster}$
            Move the k_{th} ant to the new selected feature f
            Generate random value Rand between [0,1]
          End while
        Else
          Add the $current_{cluster}$ to the $Visited_{cluster}$
          Remove the $current_{cluster}$ from the $unvisited_{cluster}$
        End if
      End while
    End for
    Update Ant_{score} for each ant.
    Calculate $\Delta\tau$ based on Ant_{score} for all features.
    $\tau[t+1] = (1-\rho) * \tau[t] + \Delta\tau[j] j = 1 \dots n$
  End for
  Sort the features by decreasing the order of their pheromones
  Select top $p * m$ features with the highest pheromone for creating S subset
End algorithm

5. Experiments

5.1. Datasets

Experiments were conducted on 10 multi-label datasets to assess the effectiveness of the proposed method. These datasets span various applications and have been widely utilized in multi-label learning tasks

Table 4
Description of multi-label datasets.

Dataset	Patterns	Features	Labels	LCardinality	LDensity	Domain
Arts	5000	462	26	1.636	0.063	Text
Business	5000	438	30	1.470	0.074	Text
CAL500	502	68	174	26.044	0.150	Music
Computer	5000	681	33	1.507	0.046	Text
Education	5000	550	33	1.461	0.044	Text
Emotions	593	72	6	1.869	0.311	Music
Entertainment	5000	640	21	1.420	0.068	Text
Health	5000	612	32	1.662	0.052	Text
Scene	2407	294	6	1.074	0.179	Image
Science	5000	743	40	1.451	0.036	Text

Table 5
Average precision of the algorithms (↑).

Dataset	MR2MLFS	MGFS	MLACO	MCLS	SCLS	MDFS	LEFMIFS
Arts	0.486164	0.477039	0.474238	0.468773	0.469163	0.481929	**0.509347**
Business	0.873005	0.8712966	**0.8741526**	0.8661106	0.8678866	0.8695826	0.862514
CAL500	**0.466253**	0.462016	0.460322	0.457103	0.463178	0.466142	0.465471
Computer	**0.625596**	0.609815	0.606406	0.587075	0.583294	0.609568	0.619541
Education	**0.549868**	0.5323558	0.5189858	0.4985708	0.5126828	0.5163138	0.540214
Emotions	0.766908	0.7856394	0.7411904	0.7173324	0.7171824	0.7622784	**0.798541**
Entertainment	**0.59368**	0.572495	0.581186	0.544426	0.564124	0.582737	0.58987
Health	**0.67825**	0.6740632	0.6699262	0.6432592	0.6600172	0.6729692	**0.67874**
Scene	**0.781313**	0.6531919	0.5924339	0.6459629	0.5562099	0.7370299	0.76584
Science	**0.490264**	0.475515	0.471783	0.444193	0.463544	0.480629	0.485412
Wilcoxon (win/tie/loss)	5/1/0	+	+	+	+	+	=

(Table 4). The datasets are freely available from the Mulan Library .[1] Each multi-label dataset $X = \{x_i, Y_i | \, 1 \leq i \leq p\}$ contains p samples, where $Y_i \in \{0,1\}^{|L|}$ defines the set of labels associated with the i^{th} sample. Label Cardinality (LCardinality) represents the average number of labels linked to each instance and is defined as follows:

$$LCardinality = \frac{1}{p}\sum_{i=1}^{p}|Y_i| \tag{19}$$

where $|Y_i|$ represents the number of labels associated with each instance. Label Density (LDensity) is defined as the label cardinality divided by the total number of labels, and is expressed as follows:

$$LDensity = \frac{1}{p}\sum_{i=1}^{I}\frac{|Y_i|}{|L|} \tag{20}$$

where $|L|$ is the size of the label set.

5.2. Baseline methods

The proposed method is evaluated against a range of state-of-the-art and baseline multi-label feature selection techniques. Details of these methods are as follows:

- **MDFS** (Zhang et al., 2019a) presents a Manifold regularized Discriminative Feature Selection for multi-label learning, that incorporates embedded manifold regularization. This method creates a low-dimensional embedding of the original feature space that accurately reflects label distributions and captures local label correlations. It also incorporates label information to account for co-occurrence relationships between label pairs.
- **MCLS** (Huang et al., 2018) Manifold-based Constraint Laplacian Score for multi-label feature selection combines laplacian scores and constraint scores, leveraging both annotated information and intrinsic data characteristics. This method applies manifold learning to convert categorical labels into numerical ones, revealing varying label importance levels for each sample.
- **MLACO** (Paniri et al., 2019) is a Multi-Label feature selection method based on the ACO. Here, the solution space is represented as a graph, and ACO is used to assign high pheromone values to prominent features. The approach applies a heuristic function to evaluate features identified by each ant. Our method similarly uses ACO to explore the solution space. However, MLACO employs a fully connected graph without addressing feature redundancy in its heuristic function. In contrast, our method constructs a graph that weights edges based on both relevancy and redundancy.
- **MGFS** (Hashemi et al., 2020b) is a Multi-label Graph-based Feature Selection algorithm that uses the PageRank algorithm to evaluate features based on their contribution to class label predictions, converting the feature set into a graph for this purpose.
- **SCLS** (Lee and Kim, 2017) uses a Scalable Criterion for large Label Set, and employs an approximation method to compute feature dependencies as well as feature relevance to label sets in a multi-label setting.
- **LEFMIFS** (Yin et al., 2024) proposes a robust multi-label approach based on Label Enhancement and Fuzzy Mutual Information Feature Selection. This method first transforms logical label matrices into continuous label distributions using a label enhancement framework that leverages local neighborhood information. It then constructs a multi-label fuzzy entropy model combining algebraic and information-theoretic views, enabling accurate estimation of feature quality.

5.3. Measures

We use the ML-kNN classifier to evaluate the comparison methods (Zhang and Zhou, 2007). This classifier finds the k nearest neighbors of each sample (here, $k = 10$) and determines target labels based on the most frequently used labels among the neighbors. Algorithm performance is assessed with several evaluation metrics, including Average Precision, F1-micro, One Error, Hamming Loss, Ranking Loss, and Coverage Error.

Average precision measures the average proportion of relevant

[1] http://mulan.sourceforge.net/datasets-mlc.html.

Table 6
F1-micro measure of the algorithms (↑).

Dataset	MR2MLFS	MGFS	MLACO	MCLS	SCLS	MDFS	LEFMIFS
Arts	**0.117415**	0.115277	0.092074	0.087974	0.09098	0.099001	0.115421
Business	0.682108	0.680408	0.680091	0.673827	0.676781	0.678519	**0.683145**
CAL500	**0.340768**	0.3313173	0.3316993	0.3211533	0.3336743	0.3353493	0.31532
Computer	**0.41928**	0.4057767	0.3975457	0.3787117	0.3875467	0.3920527	0.41567
Education	**0.286678**	0.2749538	0.2575578	0.2144528	0.2319468	0.2586448	0.235321
Emotions	0.549067	0.545359	0.50832	0.454533	0.471043	**0.567936**	0.56142
Entertainment	0.262322	0.2492585	0.2498485	0.2015085	0.2345085	0.262306	**0.314752**
Health	0.47077	0.461072	**0.475028**	0.424794	0.44598	0.462119	0.45874
Scene	0.579162	0.48349	0.375425	0.424482	0.297493	0.533391	**0.60251**
Science	**0.120981**	0.115277	0.091073	0.0507	0.075338	0.099003	0.101452
Wilcoxon (win/tie/loss)	3/2/1	+	=	+	+	=	–

Table 7
One error of the algorithms (↓).

Dataset	MR2MLFS	MGFS	MLACO	MCLS	SCLS	MDFS	LEFMIFS
Arts	0.656413	0.6662698	0.6689468	0.6883408	0.6839008	0.6751538	**0.651247**
Business	0.127758	0.127357	0.129288	0.134218	0.133713	0.131883	**0.124214**
CAL500	0.116939	0.119919	0.119794	0.12151	**0.116289**	0.117864	0.1191
Computer	**0.43549**	0.4520447	0.4468947	0.4714017	0.4580397	0.4535307	0.54201
Education	**0.569371**	0.5698822	0.5767022	0.5994652	0.5737002	0.5897352	0.570412
Emotions	0.316348	0.3193817	0.3200437	0.3491147	0.3411717	**0.3133307**	0.315412
Entertainment	**0.574083**	0.589112	0.584612	0.615731	0.607231	0.584023	0.59874
Health	**0.39502**	0.414128	0.404261	0.451951	0.434278	0.399777	**0.395142**
Scene	**0.260467**	0.279112	0.292773	0.279228	0.308124	0.274035	0.27601
Science	**0.652567**	0.665037	0.676714	0.715389	0.701183	0.67686	0.68744
Wilcoxon (win/tie/loss)	6/0/0	+	+	+	+	+	+

Table 8
Hamming loss of the algorithms (↓).

Dataset	MR2MLFS	MGFS	MLACO	MCLS	SCLS	MDFS	LEFMIFS
Arts	0.061487	0.06191	0.062625	0.062567	0.063161	0.062653	**0.05984**
Business	**0.028286**	0.0287457	0.0286967	0.0292047	0.0292457	0.0289467	0.02685
CAL500	0.140706	0.1423739	0.1425329	0.1443459	**0.1409919**	0.1415859	0.142347
Computer	**0.038858**	0.0393958	0.0398298	0.0415488	0.0409298	0.0410878	0.03924
Education	0.040851	0.041217	0.042782	0.04405	0.042363	0.041639	**0.03987**
Emotions	**0.227442**	0.243947	0.262014	0.28043	0.268565	0.233397	0.254785
Entertainment	**0.060699**	0.0615865	0.0620545	0.0647185	0.0631215	0.0624795	0.061217
Health	**0.042839**	0.0432554	0.0442644	0.0460174	0.0450434	0.0434544	0.04421
Scene	**0.127686**	0.1499615	0.1733845	0.1543385	0.1675245	0.1468315	0.14214
Science	0.034264	0.0350687	0.0349227	0.0356627	0.0352017	**0.0342117**	0.34331
Wilcoxon (win/tie/loss)	6/0/0	+	+	+	+	+	+

Table 9
Ranking loss of the algorithms (↓).

Dataset	MR2MLFS	MGFS	MLACO	MCLS	SCLS	MDFS	LEFMIFS
Arts	0.176206	**0.17551**	0.187637	0.19551	0.186292	0.18622	0.17885
Business	**0.044884**	0.0452539	0.0460569	0.0482959	0.0479589	0.0468129	0.046514
CAL500	0.188227	0.1922419	0.1925589	0.1957099	0.1879599	0.1894069	**0.183541**
Computer	**0.107031**	0.112836	0.113634	0.129912	0.130313	0.11991	0.11421
Education	**0.099438**	0.103078	0.10923	0.120232	0.11405	0.104869	**0.099871**
Emotions	**0.211804**	0.2254649	0.2262899	0.2653509	0.2386149	0.2156419	0.23541
Entertainment	**0.121603**	0.123491	0.124005	0.128748	0.126638	0.12373	0.12145
Health	**0.062868**	0.0651471	0.0646481	0.0713681	0.0683761	0.0656461	0.063654
Scene	**0.155442**	0.280436	0.377214	0.314038	0.437091	0.209869	0.21141
Science	0.137416	0.1387558	0.1383418	0.1417868	0.1399438	**0.1370648**	**0.138887**
Wilcoxon (win/tie/loss)	4/2/0	=	+	+	+	+	=

labels ranked above a given label and is defined as follows:

$$AP = \frac{1}{p}\sum_{i=1}^{I}\frac{1}{|y_i|}\sum_{\gamma\in y_i}\frac{|\gamma' \in y_i : r_i(\gamma' \le r_i(\gamma))|}{r_i(\gamma)} \tag{21}$$

where y_i represents the predictive labels and $\gamma_i(l)$ represents the rank of label $l \in L$ predicted by the learner for x_i, and p denoting the number of instances.

One Error counts the instances where the highest-confidence label is irrelevant, and it is calculated as follows:

$$OE = \frac{1}{p}\sum_{i=1}^{I}[[argmaxf(x_i, y)] \notin Y_i]y \in Y \tag{22}$$

Hamming loss represents the average discrepancy between true and predicted labels, calculated as:

$$HL = \frac{1}{I}\sum_{i=1}^{I}\frac{y_i' \oplus y_i}{q} \tag{23}$$

where $\oplus$ denotes the *XOR* operation, q and y_i' denote the size of possible labels and ground-truth labels respectively.

Coverage Error measures the average number of predicted labels required to cover all correct labels:

$$CE = \frac{1}{I}\sum_{i=1}^{I} max\ rank(\lambda) - 1\lambda \in y_i \tag{24}$$

where I is the number of instances, and $rank(\lambda)$ v the rank list of λ by likelihood (where, if $\lambda_1 > \lambda_2$, then the $rank(\lambda_1) < rank(\lambda_2)$).

Ranking Loss quantifies the instances where relevant labels are ranked lower than irrelevant labels:

$$RL = \frac{1}{I}\sum_{i=1}^{I}\frac{1}{|y_i||\bar{y}_i|}|(\lambda_1,\lambda_2)|\lambda_1 \le \lambda_2, (\lambda_1,\lambda_2) \in y_i{*}\bar{y}_i \tag{25}$$

where λ_j represent a real-valued correlation between x_i and each label $l_i \in L$, and $\bar{y}_i$ represent the complement set of y_i.

F1-micro calculates the classifier's accuracy by computing the F-measure across all predictions.

$$F1_{micro} = \frac{2\sum_{i=1}^{I}\sum_{j=1}^{l} y_{ji}h_{ji}}{\sum_{i=1}^{I}\sum_{j=1}^{l} y_{ji} + \sum_{i=1}^{I}\sum_{j=1}^{l} h_{ji}} \tag{26}$$

A lower hamming loss, one error, ranking loss, and coverage indicate improved performance, whereas a higher f1-micro and average precision signify enhanced results.

5.4. Parameter setting

The proposed method involves some tunable parameters that must be set prior to running the algorithm. The values of these parameters are determined through performing some sensitive analysis including the maximum number of iterations (NC), initial pheromone values (C_i), and the number of ants (NA) which are set to $NC = 70$, $C_i = 0.5$, and $NA = 50$, respectively. The values of parameters α and γ is set to $\alpha = 0.9$ and $\gamma = 0.1$. These parameters are employed to balance relevance and redundancy. Other parameters are configured as $= 0.2$, $q_0 = 0.6$, $r = 0.4$, and $\beta = 1$.

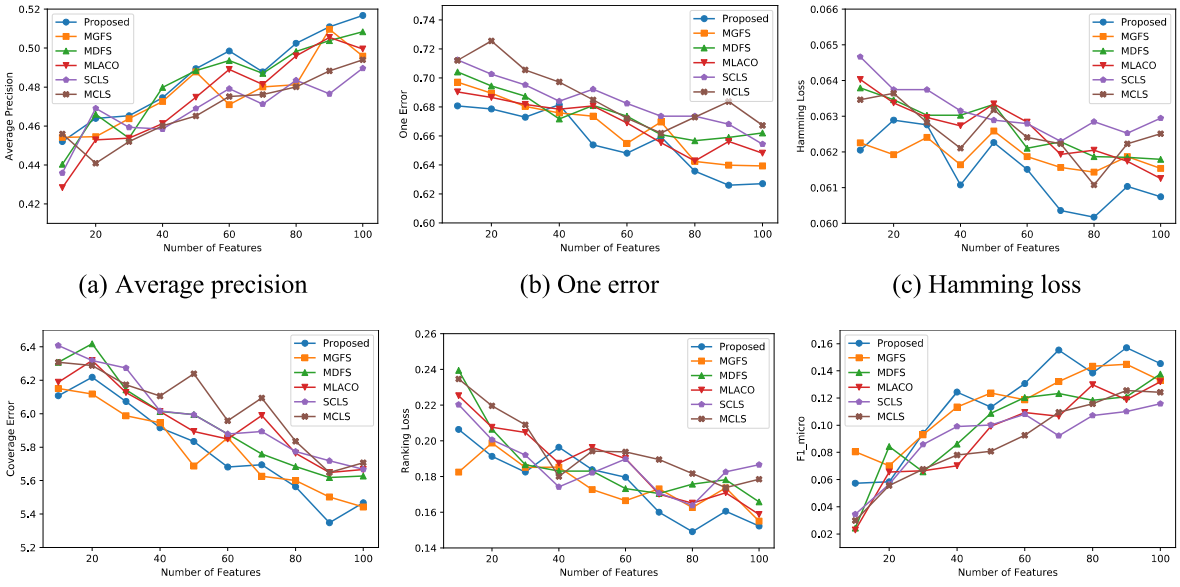
5.5. Results

The average results of all algorithms over 20 independent runs are reported. In each experiment, 70 % of the data is randomly designated for training while the remaining 30 % is utilized for testing. The algorithms were implemented using Python programming language on the Microsoft Windows 10 operating system utilizing an Intel Xeon E5-2650 v3 CPU with 64 GB of RAM.

5.5.1. Performance comparison

In these experiments, each method identifies the top m representative features, after which the dataset is modified to include only these selected features. The ML-kNN classifier is then applied to this modified dataset to evaluate the quality of the selected features. The choice of m is critical, as selecting fewer features may risk excluding relevant information. Thus, m was varied across $[10, 20, 30, \ldots, 100]$ to observe its impact. Tables 5–9 present a comparative analysis of the proposed method (MR2MLFS) against baseline methods. In each row, the best-performing value is bolded, In addition, the sign (↓) in the caption, indicates that "the smaller value is better" and (↑) indicates "the higher value is better". Tables 5 and 6 show the Average precision and F1-micro results, respectively.

The outcomes demonstrate that in most scenarios, the proposed method (MR2MLFS) outperforms the compared algorithms across multiple datasets. As shown in Table 5, MR2MLFS achieves the highest Average Precision on 7 out of 10 datasets, and ranks second-best on Business and Emotions datasets, where it is narrowly surpassed by MLACO and LEFMIFS, respectively. These results highlight the effectiveness and consistency of MR2MLFS across diverse domains. Notably, graph-based swarm intelligence methods—MR2MLFS, MLACO, and MGFS—consistently outperform manifold regularization-based approaches such as MCLS and MDFS in terms of Average Precision. This can be attributed to several factors: (1) Manifold regularization-based methods transform the original multi-label structure into Euclidean space, potentially losing critical structural dependencies between labels. (2) These methods typically rely on optimization-based objectives that



(a) Average precision (b) One error (c) Hamming loss

(d) Coverage (e) Ranking loss (f) F1-micro

Fig. 6. The performance of the algorithms as a function of the size of the optimal feature set when applied to the *Art* dataset.

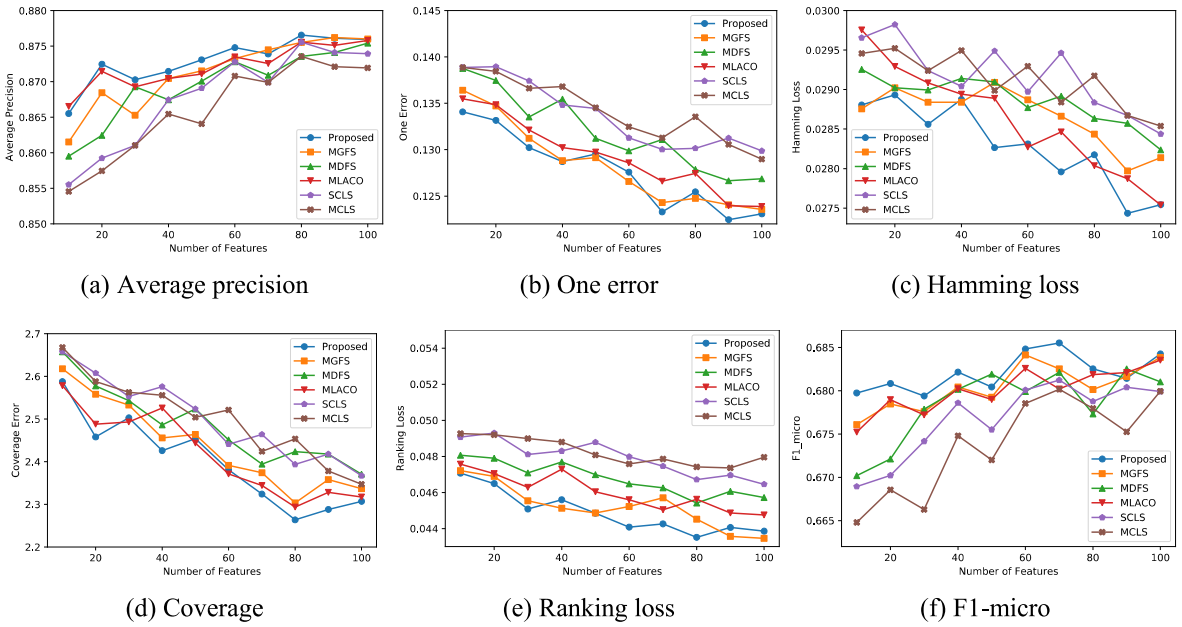


Fig. 7. The performance of the algorithms as a function of the size of the optimal feature set when applied to the *Business* dataset.

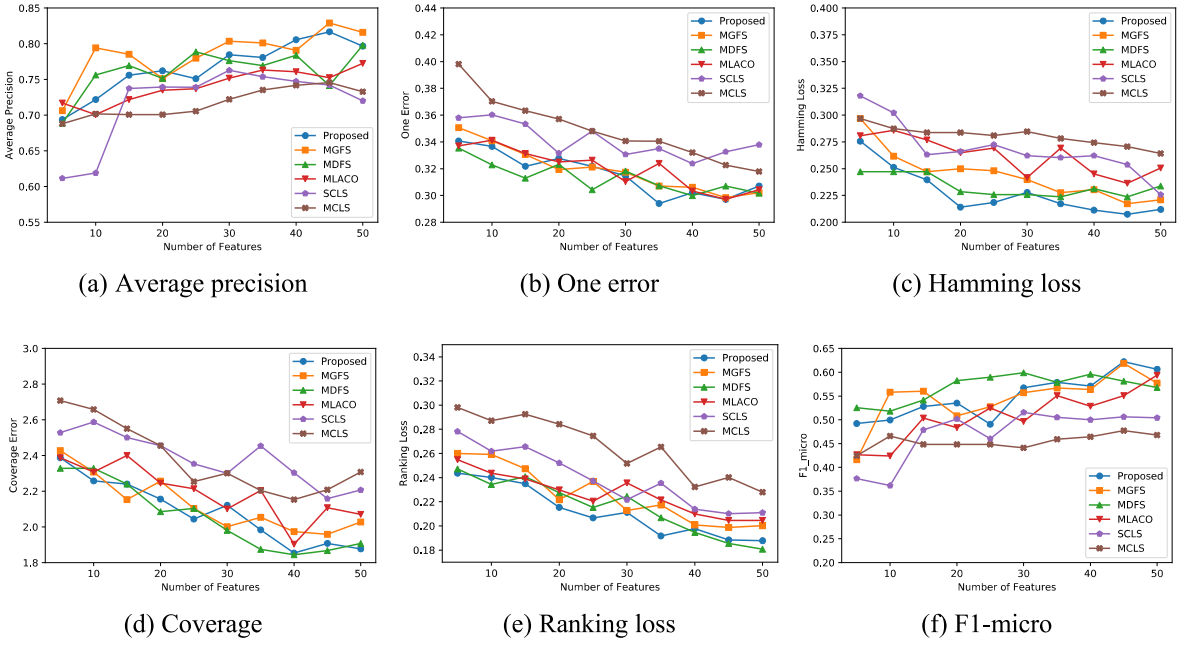


Fig. 8. The performance of the algorithms as a function of the size of the optimal feature set when applied to the *Emotions* dataset.

employ gradient descent or convex solvers, which are often susceptible to local optima, especially in high-dimensional or noisy spaces. In contrast, swarm intelligence-based approaches, particularly MR2MLFS, incorporate global and local search mechanisms to explore the solution space more effectively. Specifically, MR2MLFS leverages a multi-layered graph representation to model feature relevance and redundancy, coupled with Ant Colony Optimization (ACO) to guide the search toward highly informative and non-redundant feature subsets. The probabilistic nature of ACO and its pheromone-guided search strategy enhances exploration and mitigates premature convergence.

Table 6 presents the comparison of the proposed method with baseline algorithms in terms of the F1-micro measure, a key metric that reflects the balance between precision and recall in multi-label classification. The proposed method (MR2MLFS) delivers the best or near-best performance on the majority of datasets, including Arts, Business, CAL500, Computer, Education, Health, and Scene. While MDFS achieves marginally better results on the Emotions and Entertainment datasets, MR2MLFS remains highly competitive even in those cases, ranking as a close second. Importantly, MR2MLFS outperforms recent methods like LEFMIFS on 6 out of 10 datasets in terms of F1-micro, demonstrating its strong generalizability across various data domains. In particular, the multi-layered graph modeling and adaptive exploration of MR2MLFS enable it to select discriminative yet non-redundant features, which is critical for optimizing both precision and recall across labels. The Wilcoxon test result (win/tie/loss = 3/2/1) reinforces the statistical advantage of MR2MLFS in this metric, confirming its reliable performance in real-world multi-label learning tasks.

Tables 7–9 present a comparison of various error metrics, specifically

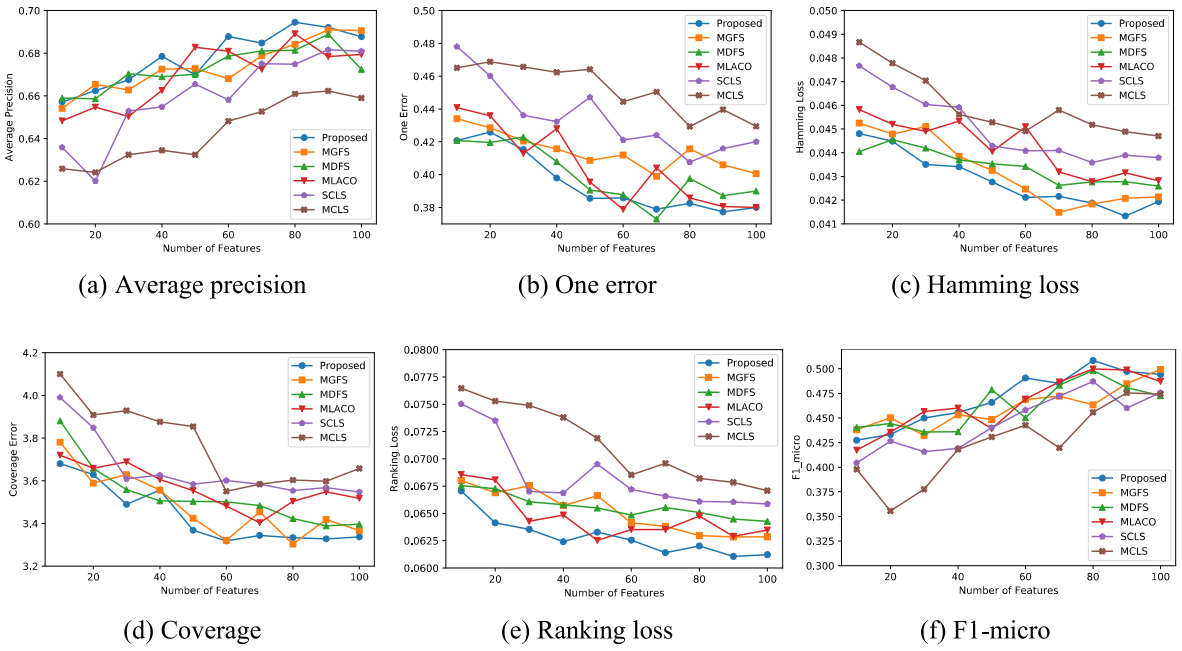


Fig. 9. The performance of the algorithms as a function of the size of the optimal feature set when applied to the *Health* dataset.

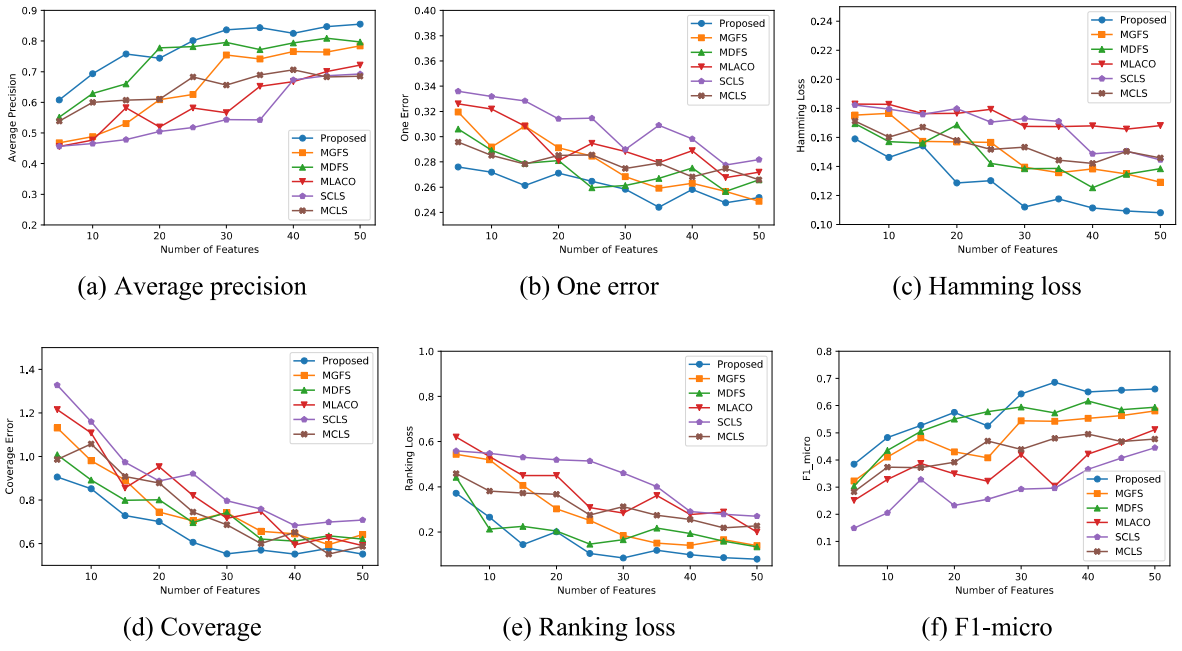


Fig. 10. The performance of the algorithms as a function of the size of the optimal feature set when applied to the *Scene* dataset.

one error, Hamming loss, and ranking loss, where lower values signify better performance.

Tables 7 and 8 present the comparative evaluation of the proposed method (MR2MLFS) and several baseline algorithms using One-Error and Hamming Loss metrics, where lower values indicate superior classification performance. The results consistently demonstrate that MR2MLFS achieves the lowest error rates on the majority of datasets, confirming its robustness in multi-label classification. Specifically, in Table 7, MR2MLFS obtains the lowest One-Error on 6 out of 10 datasets, including Arts, Business, CAL500, Computer, Scene, and Science. The Wilcoxon signed-rank test result (6/0/0) reinforces the statistical superiority of the proposed method over the competing algorithms. Interestingly, even the recently proposed LEFMIFS method—which incorporates fuzzy mutual information and label enhancement—fails to outperform MR2MLFS in most cases, although it shows slight improvements on the Arts and Business datasets. While methods such as MDFS and LEFMIFS show competitive performance on specific datasets like Emotions and Science, this can be attributed to dataset-specific characteristics such as smaller sample size or simpler label distributions which reduce the difficulty of the search space. For example, the Emotions dataset, with only 593 instances and 6 labels, does not fully exploit the advantage of MR2MLFS's layered graph-based ACO exploration strategy. In contrast, on more complex datasets such as Arts (4,498 instances), Business (11,960 instances), and Entertainment (4,957 instances), the proposed MR2MLFS method outperforms all baselines, reflecting its strength in high-dimensional, large-scale multi-label

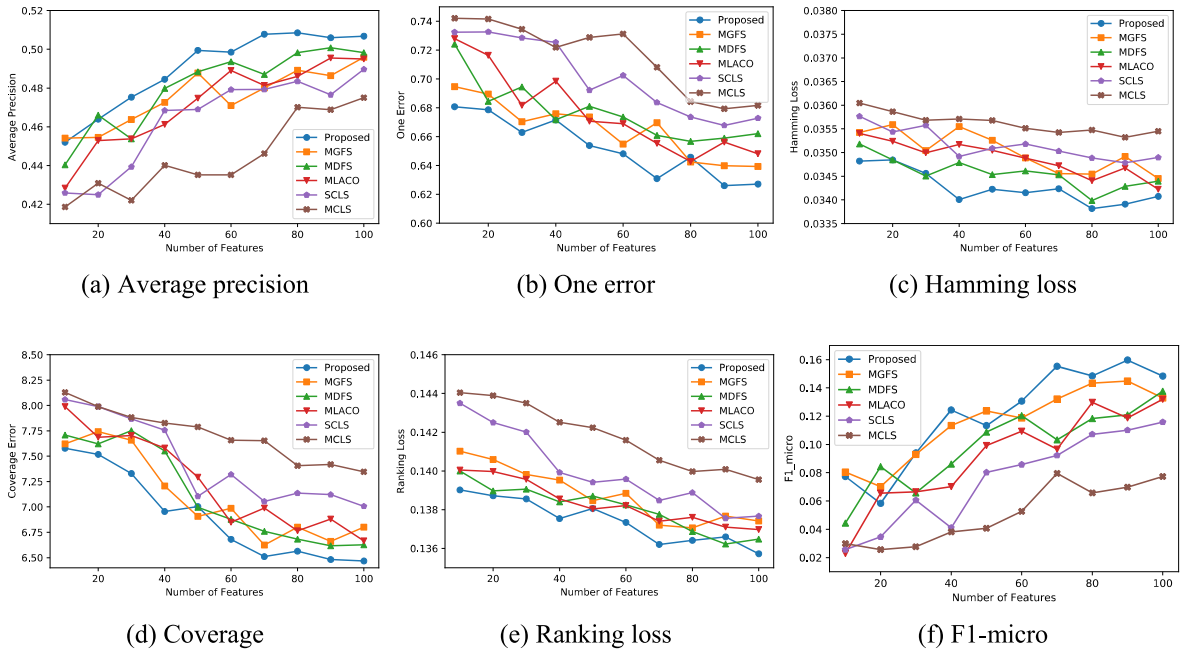


Fig. 11. The performance of the algorithms as a function of the size of the optimal feature set when applied to the *Science* dataset.

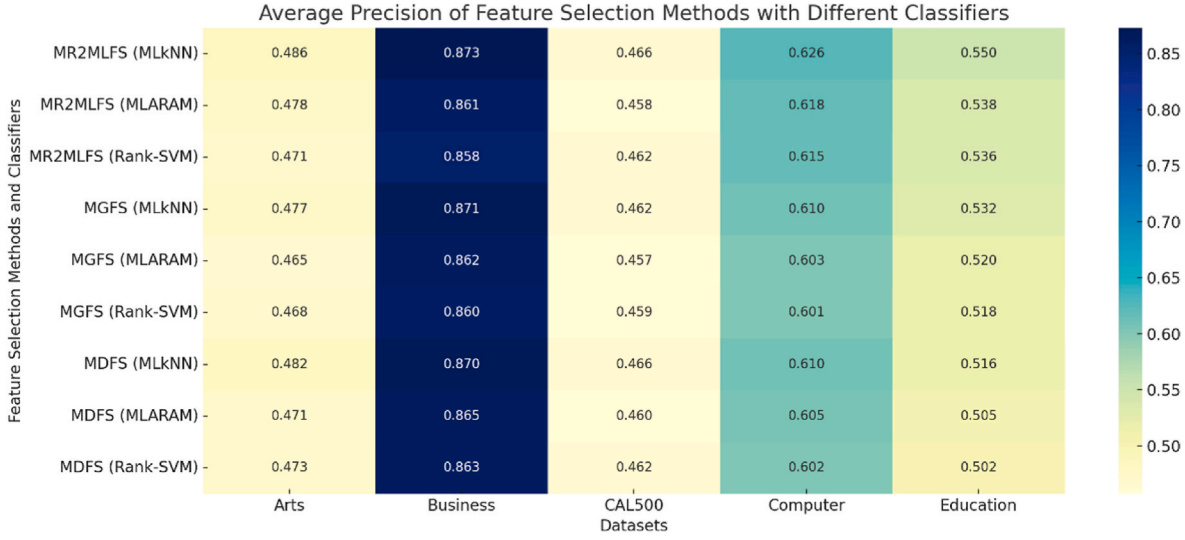


Fig. 12. Heatmap of average precision using three multi-label classifiers on five datasets.

Table 10
p-values obtained by the Friedman test based on the Average precision results.

Dataset	Vs. MDFS	Vs. MCLS	Vs. MLACO	Vs. MGFS	Vs. SCLS
Arts	0.05778	0.01141	0.00157	0.01141	0.01141
Business	0.00157	0.00157	0.01141	0.20590	0.00157
CAL500	0.20590	0.00157	0.01141	0.05778	0.20590
Computer	0.00157	0.00157	0.01141	0.00157	0.01141
Education	0.00157	0.00157	0.00157	0.01141	0.00157
Emotions	0.52709	0.00157	0.01141	0.05778	0.00157
Entertainment	0.01141	0.00157	0.01141	0.00157	0.00157
Health	0.20590	0.00157	0.05778	0.20590	0.00157
Scene	0.01141	0.00157	0.00157	0.00157	0.00157
Science	0.01141	0.00157	0.00157	0.01141	0.00157

Table 11
p-values obtained by the Friedman test based on the One error results.

Dataset	Vs. MDFS	Vs. MCLS	Vs. MLACO	Vs. MGFS	Vs. SCLS
Arts	0.01141	0.00157	0.05778	0.01141	0.00157
Business	0.00157	0.00157	0.00157	0.20590	0.00157
CAL500	0.01141	0.00157	0.00157	0.00157	0.09558
Computer	0.00157	0.00157	0.01141	0.00157	0.00157
Education	0.00157	0.00157	0.05778	1.00000	0.52709
Emotions	0.20590	0.00157	1.00000	0.20590	0.00157
Entertainment	0.01141	0.00157	0.01141	0.00157	0.00157
Health	0.09558	0.00157	0.09558	0.00157	0.00157
Scene	0.01141	0.00157	0.00157	0.01141	0.00157
Science	0.00157	0.00157	0.01141	0.01141	0.00157

learning environments. The method's ability to balance exploration and exploitation through ACO, and to preserve inter-feature and inter-label dependencies via a multi-layered graph, results in more accurate and generalizable feature subsets.

Table 9 reports the performance of the proposed MR2MLFS method and several competing algorithms in terms of Ranking Loss, a key metric in multi-label classification that evaluates the average fraction of label pairs that are incorrectly ranked. Lower values denote better

Table 12
p-values obtained by the Friedman test based on Hamming loss results.

Dataset	Vs. MDFS	Vs. MCLS	Vs. MLACO	Vs. MGFS	Vs. SCLS
Arts	0.00157	0.00157	0.00157	0.05778	0.00157
Business	0.00157	0.00157	0.20590	0.05778	0.00157
CAL500	0.20590	0.00157	0.01141	0.00157	0.52709
Computer	0.01141	0.00157	0.01141	0.20590	0.00157
Education	0.05778	0.00157	0.00157	0.05778	0.00157
Emotions	0.20590	0.00157	0.00157	0.00157	0.00157
Entertainment	0.00157	0.00157	0.01141	0.05778	0.00157
Health	0.01141	0.00157	0.00157	0.05778	0.00157
Scene	0.00157	0.00157	0.00157	0.00157	0.00157
Science	0.01141	0.00157	0.00157	0.00157	0.00157

Table 13
p-values obtained by the Friedman test based on Coverage error results.

Dataset	Vs. MDFS	Vs. MCLS	Vs. MLACO	Vs. MGFS	Vs. SCLS
Arts	0.00157	0.00157	0.00157	1.00000	0.00157
Business	0.00157	0.00157	0.52709	0.00157	0.00157
CAL500	0.31731	0.00157	0.00157	0.00157	0.05778
Computer	0.00157	0.00157	0.00157	0.00157	0.00157
Education	0.00157	0.00157	0.20590	0.20590	0.00157
Emotions	0.31731	0.00157	0.01963	0.05778	0.00157
Entertainment	0.52709	0.00157	0.09558	0.01141	0.00157
Health	0.01141	0.00157	0.00157	0.09558	0.00157
Scene	0.00157	0.00157	0.00157	0.00157	0.00157
Science	0.01141	0.00157	0.00157	0.01141	0.00157

Table 14
p-values obtained by the Friedman test based on Ranking loss results.

Dataset	Vs. MDFS	Vs. MCLS	Vs. MLACO	Vs. MGFS	Vs. SCLS
Arts	0.20590	0.01141	0.01141	0.52709	0.05778
Business	0.00157	0.00157	0.00157	0.20590	0.00157
CAL500	0.01963	0.00157	0.00157	0.00157	0.20590
Computer	0.00157	0.00157	0.01141	0.05778	0.00157
Education	0.01141	0.00157	0.01141	0.01141	0.01141
Emotions	0.52709	0.00157	0.00157	0.00157	0.00157
Entertainment	0.01141	0.00157	0.01141	0.01141	0.00157
Health	0.00157	0.00157	0.05778	0.00157	0.00157
Scene	0.01141	0.00157	0.00157	0.00157	0.00157
Science	0.01141	0.00157	0.01141	0.00157	0.00157

Table 15
p-values obtained by the Friedman test based on the F1-micro results.

Dataset	Vs. MDFS	Vs. MCLS	Vs. MLACO	Vs. MGFS	Vs. SCLS
Arts	0.01141	0.00157	0.01141	0.52709	0.00157
Business	0.05778	0.00157	0.01141	0.01141	0.00157
CAL500	0.20590	0.00157	0.01141	0.01141	0.01141
Computer	0.00157	0.00157	0.00157	0.00157	0.00157
Education	0.00157	0.00157	0.00157	0.01141	0.00157
Emotions	0.20590	0.00157	0.01141	0.20590	0.00157
Entertainment	0.52709	0.00157	0.05778	0.01141	0.00157
Health	0.01141	0.00157	1.00000	0.20590	0.00157
Scene	0.01141	0.00157	0.00157	0.00157	0.00157
Science	0.01141	0.00157	0.01141	0.20590	0.00157

performance. The results show that MR2MLFS achieves the lowest ranking loss on 6 out of 10 datasets, including Business, Computer, Education, Emotions, Health, and Entertainment. On the Arts and CAL500 datasets, MR2MLFS remains highly competitive, closely trailing the best-performing methods. For instance, on Arts, it achieves a value of 0.1762, which is nearly equivalent to MGFS (0.1755), and still better than all remaining baselines. Similarly, on CAL500, its ranking loss is marginally higher than LEFMIFS (0.1835 vs. 0.1882), showing only a slight deviation from the best. These results confirm the method's robust ranking performance, especially on larger and more complex datasets, where maintaining proper label ranking is more challenging. MR2MLFS effectively balances between maximizing relevance and minimizing redundancy through a multi-layered graph structure, which preserves important feature-label relationships. Its adaptive search strategy using ACO ensures exploration across diverse regions of the solution space, allowing the model to avoid premature convergence and obtain better feature subsets for precise label ranking.

Table 16
The results of the Friedman test (win/tie/loss) in terms of different evaluation measures. This table shows the number of win (p-value$<$0.05), tie (p-value $=$ 0.05), and loss (p-value$>$0.05) of the proposed method against the other methods.

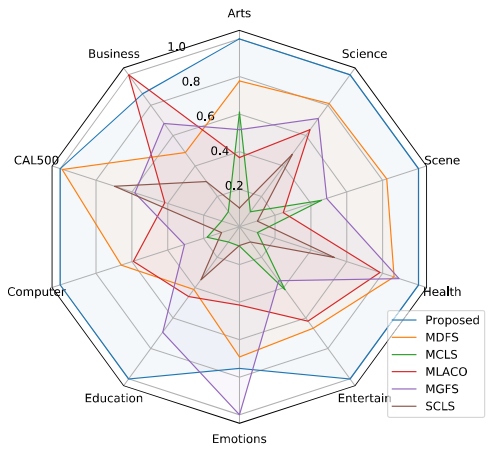
Evaluation criteria	Vs. MDFS	Vs. MCLS	Vs. MLACO	Vs. MGFS	Vs. SCLS
Average precision	6/3/1	10/0/0	9/1/0	6/4/0	9/1/0
One error	8/2/0	10/0/0	6/3/1	7/2/1	8/1/1
Hamming loss	7/3/0	10/0/0	9/1/0	4/6/0	9/0/1
Coverage error	7/2/1	10/0/0	7/2/1	6/3/1	9/1/0
Ranking loss	8/1/1	10/0/0	9/1/0	7/2/1	8/2/0
F1-micro	6/3/1	10/0/0	8/1/1	6/3/1	10/0/0
Total	42/14/4	60/0/0	48/9/3	36/20/4	53/5/2

To further assess predictive accuracy, we conducted multiple experiments to analyze the classification performance of each method as the size of the selected feature set changed. Figs. 6–11 illustrate results for various datasets, including *Art*, *Business*, *Emotions*, *Health*, *Scene*, and *Science*, with similar trends observed for other datasets. These datasets cover diverse domains, including text, music, and image data, to demonstrate the methods' adaptability across different conditions. In these graphs, the vertical axis represents classification performance based on one of the evaluation metrics, while the horizontal axis reflects the feature set size, ranging from 10 to 100. As anticipated, classification performance generally improves as the feature set size increases, as each feature contributes valuable predictive information that enhances accuracy. Overall, the results underscore that multi-label classifiers benefit significantly from feature selection. Our proposed method consistently demonstrates superior performance compared to baseline approaches. Furthermore, it delivers strong learning outcomes across most datasets, even when selecting a limited number of features, underscoring its practical effectiveness in various applications.

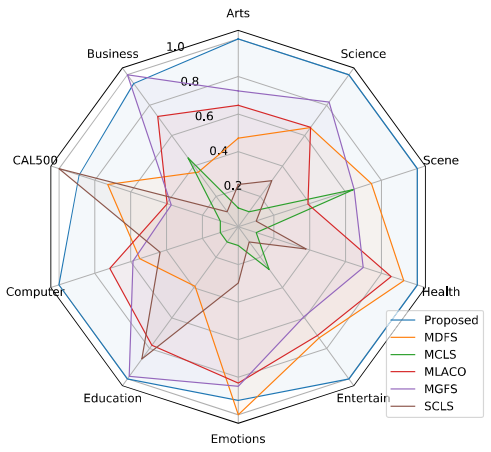
MLACO and MR2MLFS (the proposed method) both utilize the ACO algorithm to search the solution space. The results show that these two methods consistently outperform others, largely due to ACO's balanced exploration and exploitation mechanisms. MR2MLFS outperforms MLACO due to their different strategies in choosing relevant and non-redundant features. MLACO uses a fully connected graph and a heuristic metric in its search process, while the proposed MR2MLFS uses a multi-layered graph and mutual information-based ranking for a more targeted feature selection process. Similarly, MGFS and MR2MLFS share a graph-based feature-ranking foundation, with MGFS relying on the PageRank algorithm to rank features based on relevance and redundancy. The proposed method, however, incorporates ACO on a multi-layered, clustered graph to further prioritize prominent features.

5.5.2. Comparision of various multi-label classifiers

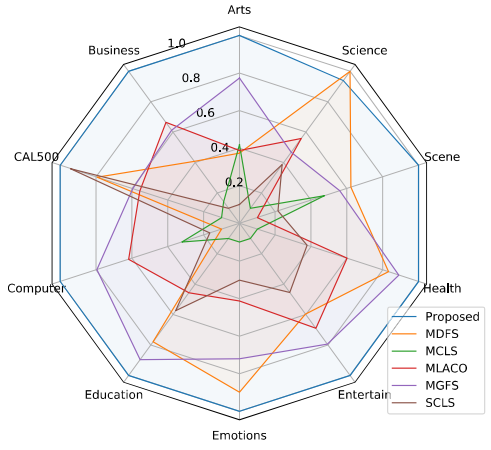
To evaluate the robustness and generalizability of the proposed feature selection method (MR2MLFS), we compared its average precision performance across three widely used multi-label classifiers (e.g. MLkNN, MLARAM, and Rank-SVM) over five representative datasets (Fig. 12). The results are also compared with MGFS and MDFS to highlight the superiority of MR2MLFS in most cases. As illustrated in the heatmap, MR2MLFS consistently achieves higher or comparable precision scores across various classifiers and datasets. Particularly in the *Business* and *Computer* datasets, MR2MLFS outperforms MGFS and MDFS in all classifiers, demonstrating its stability and effectiveness. Moreover, the minor variance across classifiers confirms the method's adaptability to different learning models, thereby addressing concerns about classifier dependency.



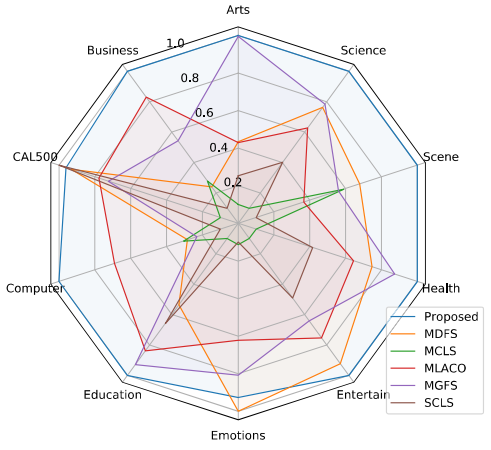
(a) Average Precision



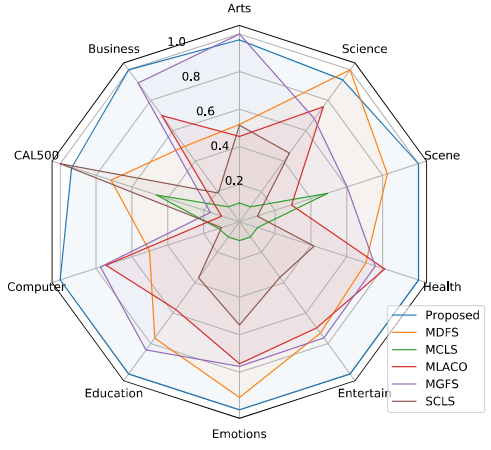
(b) One error



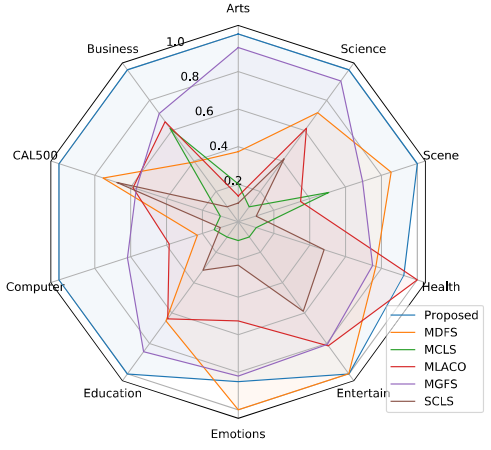
(c) Hamming Loss



(d) Coverage error



(e) Ranking Loss



(f) F1_micro

Fig. 13. The scalability analysis of the multi-label feature selection methods in terms of (a) Average precision, (b) One error, (c) Hamming loss, (d) Coverage error, (e) Ranking loss, and (f) F-micro, evaluation measures.

5.5.3. *Statistical test*

The Wilcoxon and Friedman tests were utilized to demonstrate the statistical significance of the obtained results. They are non-parametric methods that do not make any assumptions about the distribution of the data. The Wilcoxon test assesses whether the medians of paired data differ significantly, while the Friedman test compares rankings across multiple methods. Both tests calculate a p-value for paired or matched data, interpreted according to a chosen alpha significance level. In Tables 5–9, the final row presents Wilcoxon test results, where a plus sign (+) indicates that the proposed method outperforms the comparison method, an equal sign (≈) denotes comparable performance, and a minus sign (−) shows that the other method outperformed ours.

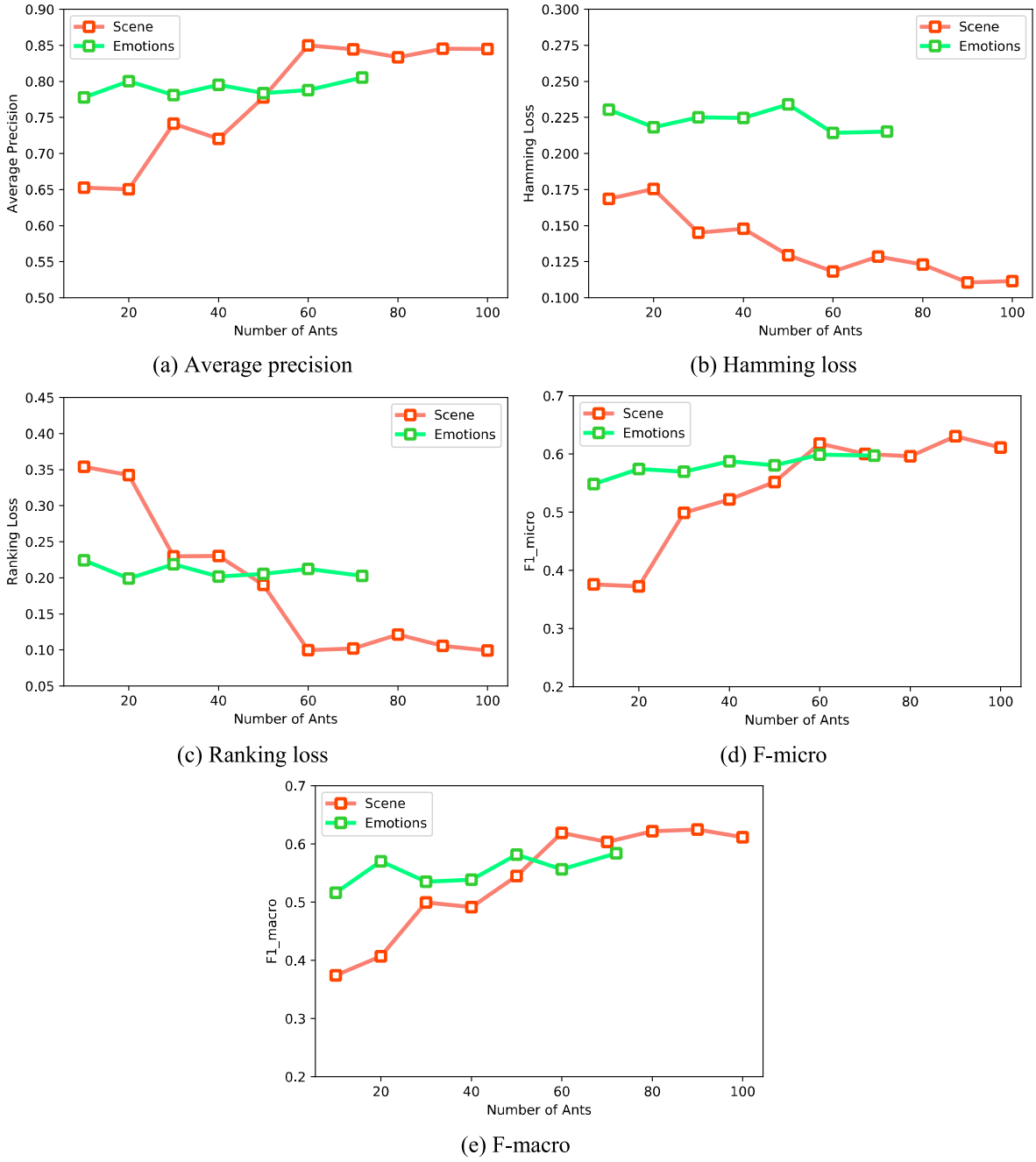


Fig. 14. The number of ants parameter analysis of the proposed method in terms of (a) Average precision, (b) Hamming loss, (c) Ranking loss, (d) F-micro, and (e) F-macro evaluation measures.

The Friedman test results are shown in Tables 10–15, with p-values for each comparison. These tables show the p-values obtained from the Friedman test and compare the proposed method against the other methods. Here if the p-value is lower than the 0.05 threshold, it shows that the proposed method statistically performed better than the other method. It can be observed from the results that in most cases, the p-values are less than $\alpha = 0.05$, indicating the rejection of the null hypothesis, and thus indicating that the proposed method outperforms the others. Moreover, similar results were reported for the other measures. For example, from Tables 11 and it can be seen that in only 8 out of 50 cases, the p-value was higher than $\alpha = 0.05$. This value for the other measures such as Hamming loss (Table 12), Coverage error (Table 13), Ranking loss (Table 14), and F1-micro (Table 15), with p-values exceeding 0.05 in only 6, 10, 8, and 9 out of 50 cases, respectively.

Table 16 presents an overall summary of the statistical comparisons, highlighting the performance of the proposed method against the baseline algorithms:

1. The proposed method consistently outperforms MCLS and SCLS across all evaluation metrics.
2. It demonstrates relatively strong performance against MLACO in terms of One Error and Coverage Error, while achieving statistically significant superiority over MLACO in the other four metrics.
3. Compared to MDFS, the proposed method performs similarly in Average Precision. However, it is statistically superior to MDFS in all other metrics, except F1-micro, where no significant difference is observed.

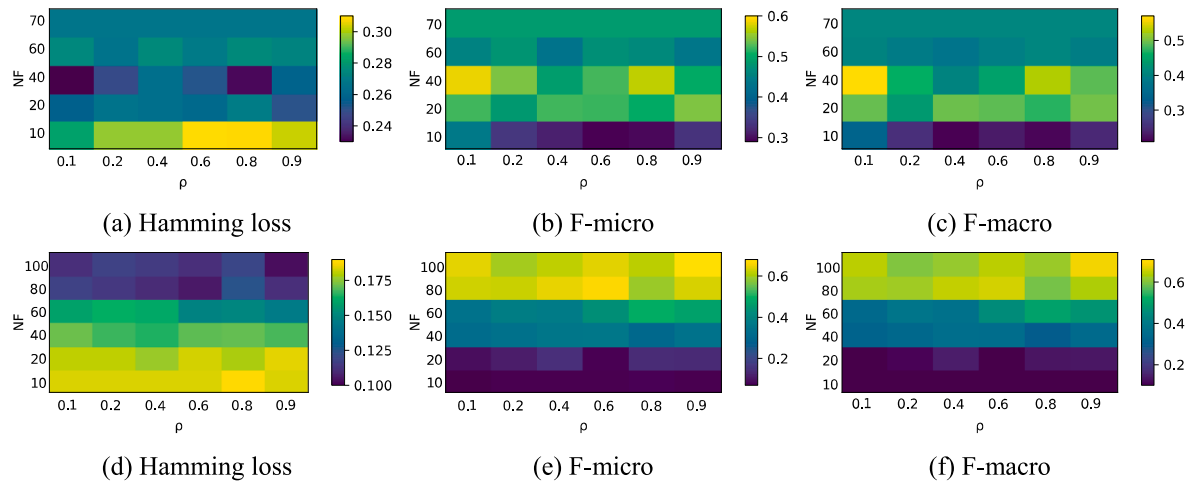


Fig. 15. Analysis of the *Rho* and *NF* parameter of the proposed method on *Emotions* dataset in terms of (a) Hamming loss, (b) F-micro, and (c) F-macro; and on *Scene* dataset in terms of (d) Hamming loss, (e) F-micro, and (f) F-macro evaluation measures.

4. Against MGFS, the proposed method achieves nearly equivalent results in Hamming Loss and relatively strong performance in Average Precision, with statistically significant improvements in the remaining three criteria.

5.5.4. Stability analysis

The stability analysis evaluates the consistency of the proposed algorithm across various evaluation metrics, with results depicted in Spider diagrams. To create these, classification performance values were normalized to the [0.1, 1] range and averaged over all feature subsets. In Fig. 13, the stability indicator for the proposed method is represented by the blue line, while other methods are denoted in violet, orange, red, brown, and green.

The Spider diagram (Fig. 13) indicates that the proposed method yields a near-regular decagonal shape, signifying consistent performance across datasets. Notable differences from other methods can be seen. For instance, Fig. 13(a) demonstrates that the proposed method achieves higher stability values for Average Precision across most datasets, except Emotions and Business. Similarly, for the Hamming Loss (Fig. 13(c)), MR2MLFS outperforms the other methods on all datasets except Science.

5.5.5. Sensitivity analysis

Analysis of the number of ants: In this section, the proposed model's reliability and parameter validity will be studied, in order to show the impact of the number of ants parameter. This parameter has been examined for each run in detail, and the obtained results have been reported. For this purpose, this parameter will be increased gradually, and the obtained results for five measures on *emotions* and *scene* datasets in the state of selecting 40 features (assuming the other parameters are constant) will be reported. As can be seen, the horizontal axes are the number of ants parameter values, and the vertical axes are the obtained results by the measures. In this experiment, the number of ants for the emotions dataset is selected between 10 and 72 (total number of the features), and for the scene dataset in the range 10–100. The green line in Fig. 14 shows the obtained results for the emotions dataset. According to the results, the increase of the number of ants parameter has little effect on F1-micro and Average precision results. However, in the final state where the number of ants is equal to the total number of features, the results have been improved slightly. At the beginning of the algorithm, for hamming loss and ranking loss measures have an unstable performance. However, by reaching the number of 40 ants, the algorithm becomes stable and its performance almost remains fixed.

For the F1-macro measure, by increasing the number of ants, the accuracy and performance will be increased accordingly. The results for the scene dataset have been shown by a red line in these figures. It is clear that by increasing the number of ants parameter, the final results have been improved; however, after reaching a certain number of ants, the algorithm's performance will not change drastically, and the algorithm will reach convergence. As shown in Fig. 14, for the F1-micro, F1-macro, Ranking loss, and Average precision metrics, by reaching the number of 60 ants, the results will be stable and do not change significantly; however, for the Hamming loss measure, the algorithm will reach stability at 80 number on ants.

Analysing ρ and *NF* parameters: In this part, we are going to analyze the influence of two important parameters on the proposed model, namely ρ: Pheromone decay and *NF*: the number of selected features by the ants in each iteration. The evaporation coefficient allows all pheromone amounts to decrease uniformly. It is a kind of forgetfulness that prevents convergence too fast. To make a thorough study of these parameters, an evaluation has been provided on the two mentioned datasets, and we reported the achieved results based on three metrics, including hamming loss, F-micro, and F-macro. In these figures, the horizontal axis shows ρ in the range [0.1, 0.9], and the vertical axis shows the *NF* in the range [10, 20, 40, 60, 70] for emotions, and [10, 20, 40, 60, 80, 100] for the scene dataset. Also, a bar has been considered near each graph that shows the lowest and highest values obtained on each dataset based on the corresponding measures using a spectrum of colors. In this bar, dark colors (dark blue to light blue) indicate low amounts, and light colors (yellow to orange) indicate high amounts of the metrics. According to Fig. 15(d–f), it is clear from scene dataset that changing the values of *NF* for each value of ρ, has made a significant change and improved the performance of the method in all measures. Also, the ρ parameter has little effect on the performance of this algorithm. In emotions dataset (Fig. 15(a–c)), for each obtained value of ρ, changing the NF from 10 to 70 could make a remarkable difference except for the last line, i.e. $NF = 70$ which stays unchanged with increasing ρ, and lies on the top row of the graphs.

In general, by examining these two parameters, it can be concluded that, on average, increasing the number of features selected by ants in each iteration leads to earlier and better identification of important features and increases the accuracy and efficiency of the introduced algorithm. Also, for each value of the NF parameter, increasing the value of the ρ parameter increases the speed of the pheromone evaporation rate from the ants in their movement path and causes the ants to forget the path traveled by the ants faster, and as a result, it causes a change in the convergence speed of the algorithm.

5.5.6. Discussion

In this paper, we presented a novel multi-label feature selection technique, MR2MLFS, which employs Ant Colony Optimization (ACO) on a multi-layer graph structure. The feature space is mapped onto a graph by considering the correlation between features and class labels. The proposed method groups similar features into the same clusters using a graph-based clustering approach. ACO is applied to this graph to assign higher weights to prominent features, and an MI-based metric is employed to assess feature subsets. This approach helps the algorithm select a set of high-quality features. To evaluate the effectiveness of our method and compare it with established techniques, several experiments were conducted on benchmark datasets. The proposed method offers several advantages that enhance its effectiveness in multi-label feature selection tasks. First, by constructing a multi-layered graph based on label-wise feature similarities and applying community detection, the method can capture both global and local interactions among features, which improves the semantic grouping and interpretability of selected features. Second, the integration of the ACO allows for a robust and adaptive search through the feature space, balancing exploration and exploitation in a dynamic manner. This combination leads to the selection of features that are not only relevant but also minimally redundant. Third, the modular nature of the approach makes it easily extensible to different domains and adaptable to high-dimensional datasets. Despite these strengths, the proposed method also has certain limitations. One key limitation is its computational complexity, especially during the community detection and ACO-based search stages, which may become resource-intensive when applied to very large-scale datasets. Furthermore, the performance of the method can be sensitive to the choice of parameters in the ACO process (e.g., pheromone decay rate, number of ants), which requires careful tuning. Additionally, while the Louvain algorithm is efficient and effective for community detection, it may yield different results on networks with small modularity gains, potentially affecting consistency in feature grouping. Future work will aim to address these limitations by exploring parallel or distributed implementations to enhance scalability. The results indicate that the proposed method outperformed alternative approaches. Additionally, the following findings were derived from the results:

1 The results presented in Tables 5–9 indicate that, in most scenarios, the proposed method achieves the best results. This superiority is especially pronounced on non-sparse datasets, such as Scene.
2 The datasets used in the experiments span various domains, including text, image, and audio. Our method consistently achieved excellent performance across all these domains, indicating that MR2MLFS is not biased toward any specific data type.
3 In ACO-based methods (MLACO and MR2MLFS), the ants can traverse the graph in parallel. Thus, compared to other evolutionary methods, these approaches generally require fewer computational resources.
4 Both MLACO and MR2MLFS employ ACO in their processes. However, the experimental results demonstrate that the proposed method consistently outperforms MLACO. This improvement is due to two primary reasons. First, our method considers the relationship between features and labels when mapping features to the graph, and it clusters the graph to group redundant features within the same clusters. Second, the ACO search process is modified to traverse a multi-layered graph, using an MI-based criterion that incorporates both relevance and redundancy in weighting features.
5 The experiments also revealed that the proposed method outperforms MGFS. While both MGFS and MR2MLFS apply a similar approach to map the feature space onto a graph, MGFS uses PageRank for feature ranking, which only considers graph structure. In contrast, MR2MLFS's search process incorporates both relevance and redundancy, enhancing the effectiveness of feature selection.

To further improve scalability on large-scale, high-dimensional datasets, the proposed method can be adapted in several ways. First, the ACO search process is inherently parallelisable, as each ant constructs its solution independently. This property makes it straightforward to distribute computation across multiple processing cores or machines to significantly reduce runtime. Additionally, a pre-filtering step based on simple statistical measures (e.g., variance thresholding or correlation filtering) can be applied to reduce the number of candidate features before constructing the multi-layer graph, further improving efficiency without compromising the quality of the selected subset. For very large feature spaces, the feature graph could be partitioned into smaller subgraphs, and ACO could be applied iteratively or hierarchically on these partitions. Future work will also explore hybrid strategies that combine the proposed graph-based ACO with deep learning-based feature embeddings to accelerate the search process and improve scalability in high-dimensional applications.

6. Conclusion and future works

The proposed method integrates the ACO with a multi-layered graph representation to enhance multi-label feature selection. It begins by transforming the feature space into an undirected, weighted graph, where edge weights reflect the relevance of features to the target labels. To reduce redundancy, a graph clustering technique groups similar features into communities. The extended ACO search process then operates over this multi-layered graph, with ants favouring transitions between meta-nodes (clusters) in the first layer while less frequently exploring within individual clusters. A novel mutual information-based metric is used to evaluate feature subsets by simultaneously considering their relevance to target labels and their redundancy with other features, allowing higher pheromone values to accumulate on the most informative features. The final feature set is formed by selecting those with the highest pheromone ranks. Experiments conducted on diverse benchmark datasets demonstrate that MR2MLFS outperforms state-of-the-art algorithms in most scenarios. Despite these promising results, the method has some limitations. The computational complexity involved in constructing multi-layer graphs and performing ACO-based optimization can be substantial, especially for high-dimensional or large-scale datasets. Additionally, the performance is sensitive to several ACO hyperparameters, which may require careful fine-tuning to achieve optimal results.

There are several promising directions for future research. One direction is adapting the method for streaming data to enable real-time feature selection in dynamic environments. Extending the approach to handle datasets with missing or noisy labels would further increase its practical applicability. Improving scalability through parallel or distributed implementations is another key direction, particularly for large-scale, high-dimensional problems. Exploring alternative or more advanced community detection methods may also improve the quality of the multi-layered graph representation. Additionally, explicitly modelling label dependencies within the graph structure could lead to better performance when labels are highly correlated. Finally, integrating deep learning techniques, such as graph neural networks, may offer more powerful feature selection by jointly learning feature relevance and redundancy.

CRediT authorship contribution statement

Mohammad Hatami: Writing – original draft, Visualization, Software, Resources, Investigation, Formal analysis, Data curation. **Parham Moradi:** Writing – review & editing, Validation, Supervision, Project administration, Methodology, Conceptualization. **Sadegh Sulaimany:** Supervision, Resources. **Mahdi Jalili:** Writing – review & editing, Validation, Project administration, Methodology, Conceptualization.

Funding

Mahdi Jalili is supported by Australian Research Council through projects DP240100963, LP230100439, and IM240100042.

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.

Data availability

Data will be made available on request.

References

Akhtar, M.U., Liu, J., Xie, Z., Cui, X., Liu, X., Huang, B., 2025. Multilingual entity alignment by abductive knowledge reasoning on multiple knowledge graphs. Eng. Appl. Artif. Intell. 139, 109660. https://doi.org/10.1016/j.engappai.2024.109660.

Asilian Bidgoli, A., Ebrahimpour-Komleh, H., Rahnamayan, S., 2021. Reference-point-based multi-objective optimization algorithm with opposition-based voting scheme for multi-label feature selection. Inf. Sci. 547, 1–17. https://doi.org/10.1016/j.ins.2020.08.004.

Blondel, V.D., Guillaume, J.-L., Lambiotte, R., Lefebvre, E., 2008. Fast unfolding of communities in large networks. J. Stat. Mech. Theor. Exp. 2008, P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008.

Che, X., Chen, D., Mi, J., 2020. A novel approach for learning label correlation with application to feature selection of multi-label data. Inf. Sci. 512, 795–812. https://doi.org/10.1016/j.ins.2019.10.022.

Dadaneh, B.Z., Markid, H.Y., Zakerolhosseini, A., 2016. Unsupervised probabilistic feature selection using ant colony optimization. Expert Syst. Appl. 53, 27–42. https://doi.org/10.1016/j.eswa.2016.01.021.

Dai, J., Chen, J., Liu, Y., Hu, H., 2020. Novel multi-label feature selection via label symmetric uncertainty correlation learning and feature redundancy evaluation. Knowl. Base Syst. 207, 106342. https://doi.org/10.1016/j.knosys.2020.106342.

Dai, J., Huang, W., Zhang, C., Liu, J., 2024. Multi-label feature selection by strongly relevant label gain and label mutual aid. Pattern Recogn. 145, 109945. https://doi.org/10.1016/j.patcog.2023.109945.

Dai, J., Wang, J., 2025. Multi-label feature selection with missing features by tolerance implication granularity information and symmetric coupled discriminant weight. Pattern Recogn. 162, 111365. https://doi.org/10.1016/j.patcog.2025.111365.

Ding, J., Qian, W., Li, Y., Yang, W., Huang, J., 2024. Partial label feature selection via label disambiguation and neighborhood mutual information. Inf. Sci. 680, 121163. https://doi.org/10.1016/j.ins.2024.121163.

Dong, H., Sun, J., Sun, X., Ding, R., 2020. A many-objective feature selection for multi-label classification. Knowl. Base Syst. 208, 106456. https://doi.org/10.1016/j.knosys.2020.106456.

Fan, Y., Chen, B., Huang, W., Liu, J., Weng, W., Lan, W., 2022. Multi-label feature selection based on label correlations and feature redundancy. Knowl. Base Syst. 241, 108256. https://doi.org/10.1016/j.knosys.2022.108256.

Faraji, M., Seyedi, S.A., Akhlaghian Tab, F., Mahmoodi, R., 2024. Multi-label feature selection with global and local label correlation. Expert Syst. Appl. 246, 123198. https://doi.org/10.1016/j.eswa.2024.123198.

Ghimatgar, H., Kazemi, K., Helfroush, M.S., Aarabi, A., 2018. An improved feature selection algorithm based on graph clustering and ant colony optimization. Knowl. Base Syst. 159, 270–285. https://doi.org/10.1016/j.knosys.2018.06.025.

Hamedmoghadam, H., Jalili, M., Yu, X., 2018. An opinion formation based binary optimization approach for feature selection. Phys. Stat. Mech. Appl. 491, 142–152. https://doi.org/10.1016/j.physa.2017.08.048.

Han, Q., Zhao, Z., Hu, L., Gao, W., 2025. Enhanced multi-label feature selection considering label-specific relevant information. Expert Syst. Appl. 264, 125819. https://doi.org/10.1016/j.eswa.2024.125819.

Hao, P., Ding, W., Gao, W., He, J., 2024. Exploring view-specific label relationships for multi-view multi-label feature selection. Inf. Sci. 681, 121215. https://doi.org/10.1016/j.ins.2024.121215.

Hashemi, A., Dowlatshahi, M.B., Nezamabadi-pour, H., 2020a. MFS-MCDM: multi-label feature selection using multi-criteria decision making. Knowl. Base Syst. 206, 106365. https://doi.org/10.1016/j.knosys.2020.106365.

Hashemi, A., Dowlatshahi, M.B., Nezamabadi-pour, H., 2020b. MGFS: a multi-label graph-based feature selection algorithm via PageRank centrality. Expert Syst. Appl. 142, 113024. https://doi.org/10.1016/j.eswa.2019.113024.

Hatami, M., Mahmood, S.R., Moradi, P., 2020a. A Graph-based multi-label feature selection using ant colony optimization. 2020 10th International Symposium Ontelecommunications (IST), pp. 175–180. https://doi.org/10.1109/IST50524.2020.9345913.

Hatami, M., Mehrmohammadi, P., Moradi, P., 2020b. A multi-label feature selection based on mutual information and ant colony optimization. 2020 28th Iranian Conference on Electrical Engineering (ICEE), pp. 1–6. https://doi.org/10.1109/ICEE50131.2020.9260852.

He, Z.-F., Yang, M., Gao, Y., Liu, H.-D., Yin, Y., 2019. Joint multi-label classification and label correlations with missing labels and feature selection. Knowl. Base Syst. 163, 145–158. https://doi.org/10.1016/j.knosys.2018.08.018.

Hu, L., Li, Y., Gao, W., Zhang, P., Hu, J., 2020. Multi-label feature selection with shared common mode. Pattern Recogn., 107344 https://doi.org/10.1016/j.patcog.2020.107344.

Huang, R., Jiang, W., Sun, G., 2018. Manifold-based constraint Laplacian score for multi-label feature selection. Pattern Recognit. Lett. 112, 346–352. https://doi.org/10.1016/j.patrec.2018.08.021.

Karimi, F., Dowlatshahi, M.B., Hashemi, A., 2023. SemiACO: a semi-supervised feature selection based on ant colony optimization. Expert Syst. Appl. 214, 119130. https://doi.org/10.1016/j.eswa.2022.119130.

Kashef, S., Nezamabadi-pour, H., 2015. An advanced ACO algorithm for feature subset selection. Neurocomputing 147, 271–279. https://doi.org/10.1016/j.neucom.2014.06.067.

Kashef, S., Nezamabadi-pour, H., 2019. A label-specific multi-label feature selection algorithm based on the pareto dominance concept. Pattern Recogn. 88, 654–667. https://doi.org/10.1016/j.patcog.2018.12.020.

Kashef, S., Nezamabadi-pour, H., Nikpour, B., 2018. Multilabel feature selection: a comprehensive review and guiding experiments. WIREs Data Mining and Knowledge Discovery 8, e1240. https://doi.org/10.1002/widm.1240.

Labani, M., Moradi, P., Jalili, M., 2020. A multi-objective genetic algorithm for text feature selection using the relative discriminative criterion. Expert Syst. Appl. 149, 113276. https://doi.org/10.1016/j.eswa.2020.113276.

Lee, J., Kim, D.-W., 2013. Feature selection for multi-label classification using multivariate mutual information. Pattern Recognit. Lett. 34, 349–357. https://doi.org/10.1016/j.patrec.2012.10.005.

Lee, J., Kim, D.-W., 2017. SCLS: multi-label feature selection based on scalable criterion for large label set. Pattern Recogn. 66, 342–352. https://doi.org/10.1016/j.patcog.2017.01.014.

Lee, J., Yu, I., Park, J., Kim, D.-W., 2019. Memetic feature selection for multilabel text categorization using label frequency difference. Inf. Sci. 485, 263–280. https://doi.org/10.1016/j.ins.2019.02.021.

Li, F., Miao, D., Pedrycz, W., 2017. Granular multi-label feature selection based on mutual information. Pattern Recogn. 67, 410–423. https://doi.org/10.1016/j.patcog.2017.02.025.

Li, X., Fu, Q., Li, Q., Ding, W., Lin, F., Zheng, Z., 2023. Multi-objective binary grey wolf optimization for feature selection based on guided mutation strategy. Appl. Soft Comput. 145, 110558. https://doi.org/10.1016/j.asoc.2023.110558.

Li, Y., Hu, L., Gao, W., 2022. Label correlations variation for robust multi-label feature selection. Inf. Sci. 609, 1075–1097. https://doi.org/10.1016/j.ins.2022.07.154.

Li, Y., Li, M., Zhang, X., Ding, J., 2025a. Dynamic Q&A multi-label classification based on adaptive multi-scale feature extraction. Appl. Soft Comput. 170, 112740. https://doi.org/10.1016/j.asoc.2025.112740.

Li, Z., Li, H., Gao, W., Xie, J., Slowik, A., 2025b. Feature selection in high-dimensional classification via an adaptive multifactor evolutionary algorithm with local search. Appl. Soft Comput. 169, 112574. https://doi.org/10.1016/j.asoc.2024.112574.

Lin, Y., Hu, Q., Liu, J., Chen, J., Duan, J., 2016. Multi-label feature selection based on neighborhood mutual information. Appl. Soft Comput. 38, 244–256. https://doi.org/10.1016/j.asoc.2015.10.009.

Lv, S., Shi, S., Wang, H., Li, F., 2021. Semi-supervised multi-label feature selection with adaptive structure learning and manifold learning. Knowl. Base Syst. 214, 106757. https://doi.org/10.1016/j.knosys.2021.106757.

Ma, W., Zhou, X., Zhu, H., Li, L., Jiao, L., 2021. A two-stage hybrid ant colony optimization for high-dimensional feature selection. Pattern Recogn. 116, 107933. https://doi.org/10.1016/j.patcog.2021.107933.

Mahmood, S.R., Hatami, M., Moradi, P., 2020. A trust-based recommender system by integration of graph clustering and ant colony optimization. 2020 10th International Conference on Computer and Knowledge Engineering (ICCKE), pp. 598–604. https://doi.org/10.1109/ICCKE50421.2020.9303647.

Mehrmohammadi, P., Hatami, M., Moradi, P., 2020. A density peaks clustering method based on mutual kNN graph and shortest path. 2020 28th Iranian Conference on Electrical Engineering. ICEE), pp. 1–6. https://doi.org/10.1109/ICEE50131.2020.9260954.

Mehrmohammadi, P., Hatami, M., Moradi, P., 2021. A graph-based density peaks method by employing shortest path for data clustering. 2021 5th International Conference on Pattern Recognition and Image Analysis (IPRIA), pp. 1–8. https://doi.org/10.1109/IPRIA53572.2021.9483576.

Miao, F., Wu, Y., Yan, G., Si, X., 2025. Dynamic multi-swarm whale optimization algorithm based on elite tuning for high-dimensional feature selection classification problems. Appl. Soft Comput. 169, 112634. https://doi.org/10.1016/j.asoc.2024.112634.

Moradi, P., Rostami, M., 2015. Integration of graph clustering with ant colony optimization for feature selection. Knowl. Base Syst. 84, 144–161. https://doi.org/10.1016/j.knosys.2015.04.007.

Nemati, K., Refahi Sheikhani, A.H., Kordrostami, S., Khoshhal Roudposhti, K., 2024. New hybrid feature selection approaches based on ANN and novel sparsity norm. Journal of Electrical and Computer Engineering 2024, 7112770. https://doi.org/10.1155/2024/7112770.

Paniri, M., Dowlatshahi, M.B., Nezamabadi-pour, H., 2019. MLACO: a multi-label feature selection algorithm based on ant colony optimization. Knowl. Base Syst., 105285 https://doi.org/10.1016/j.knosys.2019.105285.

Paniri, M., Dowlatshahi, M.B., Nezamabadi-pour, H., 2021. Ant-TD: Ant colony optimization plus temporal difference reinforcement learning for multi-label feature selection. Swarm Evol. Comput. 64, 100892. https://doi.org/10.1016/j.swevo.2021.100892.

Paul, D., Jain, A., Saha, S., Mathew, J., 2021. Multi-objective PSO based online feature selection for multi-label classification. Knowl. Base Syst. 222, 106966. https://doi.org/10.1016/j.knosys.2021.106966.

Qian, W., Huang, J., Wang, Y., Shu, W., 2020. Mutual information-based label distribution feature selection for multi-label learning. Knowl. Base Syst., 105684 https://doi.org/10.1016/j.knosys.2020.105684.

Qian, W., Ye, Q., Li, Y., Huang, J., Dai, S., 2022. Relevance-based label distribution feature selection via convex optimization. Inf. Sci. 607, 322–345. https://doi.org/10.1016/j.ins.2022.05.094.

Rafie, A., Moradi, P., Ghaderzadeh, A., 2023. A multi-objective online streaming multi-label feature selection using mutual information. Expert Syst. Appl. 216, 119428. https://doi.org/10.1016/j.eswa.2022.119428.

Rahmaninia, M., Moradi, P., 2018. OSFSMI: online stream feature selection method based on mutual information. Appl. Soft Comput. 68, 733–746. https://doi.org/10.1016/j.asoc.2017.08.034.

Rao, H., Shi, X., Rodrigue, A.K., Feng, J., Xia, Y., Elhoseny, M., Yuan, X., Gu, L., 2019. Feature selection based on artificial bee colony and gradient boosting decision tree. Appl. Soft Comput. 74, 634–642. https://doi.org/10.1016/j.asoc.2018.10.036.

Salem, O.A.M., Liu, F., Chen, Y.-P.P., Hamed, A., Chen, X., 2022. Effective fuzzy joint mutual information feature selection based on uncertainty region for classification problem. Knowl. Base Syst. 257, 109885. https://doi.org/10.1016/j.knosys.2022.109885.

Sharmin, S., Shoyaib, M., Ali, A.A., Khan, M.A.H., Chae, O., 2019. Simultaneous feature selection and discretization based on mutual information. Pattern Recogn. 91, 162–174. https://doi.org/10.1016/j.patcog.2019.02.016.

Spolaôr, N., Monard, M.C., Tsoumakas, G., Lee, H.D., 2016. A systematic review of multi-label feature selection and a new method based on label construction. Neurocomputing 180, 3–15. https://doi.org/10.1016/j.neucom.2015.07.118.

Sun, Z., Zhang, J., Dai, L., Li, C., Zhou, C., Xin, J., Li, S., 2019. Mutual information based multi-label feature selection via constrained convex optimization. Neurocomputing 329, 447–456. https://doi.org/10.1016/j.neucom.2018.10.047.

Tabakhi, S., Moradi, P., 2015. Relevance–redundancy feature selection based on ant colony optimization. Pattern Recogn. 48, 2798–2811. https://doi.org/10.1016/j.patcog.2015.03.020.

Tabakhi, S., Moradi, P., Akhlaghian, F., 2014. An unsupervised feature selection algorithm based on ant colony optimization. Eng. Appl. Artif. Intell. 32, 112–123. https://doi.org/10.1016/j.engappai.2014.03.007.

Teisseyre, P., Zufferey, D., Słomka, M., 2019. Cost-sensitive classifier chains: selecting low-cost features in multi-label classification. Pattern Recogn. 86, 290–319. https://doi.org/10.1016/j.patcog.2018.09.012.

Wan, Y., Wang, M., Ye, Z., Lai, X., 2016. A feature selection method based on modified binary coded ant colony optimization algorithm. Appl. Soft Comput. 49, 248–258. https://doi.org/10.1016/j.asoc.2016.08.011.

Wang, D., Wang, L., Chen, W., Wang, H., Liang, C., 2025. Unsupervised multi-view feature selection based on weighted low-rank tensor learning and its application in multi-omics datasets. Eng. Appl. Artif. Intell. 143, 110041. https://doi.org/10.1016/j.engappai.2025.110041.

Wang, X.-h., Zhang, Y., Sun, X.-y., Wang, Y.-l., Du, C.-h., 2020. Multi-objective feature selection based on artificial bee colony: an acceleration approach with variable sample size. Appl. Soft Comput. 88, 106041. https://doi.org/10.1016/j.asoc.2019.106041.

Xiong, C., Qian, W., Wang, Y., Huang, J., 2021. Feature selection based on label distribution and fuzzy mutual information. Inf. Sci. 574, 297–319. https://doi.org/10.1016/j.ins.2021.06.005.

Yang, Y., Chen, H., Mi, Y., Luo, C., Horng, S.-J., Li, T., 2023. Multi-label feature selection based on stable label relevance and label-specific features. Inf. Sci. 648, 119525. https://doi.org/10.1016/j.ins.2023.119525.

Yin, T., Chen, H., Yuan, Z., Sang, B., Horng, S.-J., Li, T., Luo, C., 2024. LEFMIFS: label enhancement and fuzzy mutual information for robust multilabel feature selection. Eng. Appl. Artif. Intell. 133, 108108. https://doi.org/10.1016/j.engappai.2024.108108.

You, H., Wang, P., Li, Z., 2024. Feature selection for label distribution learning based on the statistical distribution of data and fuzzy mutual information. Inf. Sci. 679, 121085. https://doi.org/10.1016/j.ins.2024.121085.

Zandvakili, A., Mansouri, N., Javidi, M.M., 2024. A new feature selection algorithm based on fuzzy-pathfinder optimization. Neural Comput. Appl. https://doi.org/10.1007/s00521-024-10043-2.

Zhang, H., Qin, X., Gao, X., Zhang, S., Tian, Y., Zhang, W., 2024a. Modified salp swarm algorithm based on competition mechanism and variable shifted windows for feature selection. Soft Comput. https://doi.org/10.1007/s00500-024-09876-9.

Zhang, J., Luo, Z., Li, C., Zhou, C., Li, S., 2019a. Manifold regularized discriminative feature selection for multi-label learning. Pattern Recogn. https://doi.org/10.1016/j.patcog.2019.06.003.

Zhang, M.-L., Zhou, Z.-H., 2007. ML-KNN: a lazy learning approach to multi-label learning. Pattern Recogn. 40, 2038–2048. https://doi.org/10.1016/j.patcog.2006.12.019.

Zhang, N., Wang, A., Lu, P., Feng, T., Xu, Y., Du, G., 2025. Multi-label feature selection with feature reconstruction and label correlations. Expert Syst. Appl. 285, 127993. https://doi.org/10.1016/j.eswa.2025.127993.

Zhang, P., Liu, G., Gao, W., 2019b. Distinguishing two types of labels for multi-label feature selection. Pattern Recogn. 95, 72–82. https://doi.org/10.1016/j.patcog.2019.06.004.

Zhang, P., Liu, G., Song, J., 2023. MFSJMI: multi-Label feature selection considering join mutual information and interaction weight. Pattern Recogn. 138, 109378. https://doi.org/10.1016/j.patcog.2023.109378.

Zhang, Q., Liu, S., Wang, J., Li, Z., Wen, C.-F., 2024b. Feature selection for multi-labeled data based on label enhancement technique and mutual information. Inf. Sci. 679, 121113. https://doi.org/10.1016/j.ins.2024.121113.

Zhang, Y., Cheng, S., Shi, Y., Gong, D.-w., Zhao, X., 2019c. Cost-sensitive feature selection using two-archive multi-objective artificial bee colony algorithm. Expert Syst. Appl. 137, 46–58. https://doi.org/10.1016/j.eswa.2019.06.044.

Zhang, Y., Gong, D.-w., Sun, X.-y., Guo, Y.-n., 2017. A PSO-based multi-objective multi-label feature selection method in classification. Sci. Rep. 7, 1–12. https://doi.org/10.1038/s41598-017-00416-0.

Zhou, G., Li, R., Shang, Z., Li, X., Jia, L., 2024a. Multi-label feature selection based on minimizing feature redundancy of mutual information. Neurocomputing 607, 128392. https://doi.org/10.1016/j.neucom.2024.128392.

Zhou, X., Ma, H., Gu, J., Chen, H., Deng, W., 2022. Parameter adaptation-based ant colony optimization with dynamic hybrid mechanism. Eng. Appl. Artif. Intell. 114, 105139. https://doi.org/10.1016/j.engappai.2022.105139.

Zhou, X., Yuan, W., Gao, Q., Yang, C., 2024b. An efficient ensemble learning method based on multi-objective feature selection. Inf. Sci. 679, 121084. https://doi.org/10.1016/j.ins.2024.121084.

Zhou, Y., Zhang, W., Kang, J., Zhang, X., Wang, X., 2021. A problem-specific non-dominated sorting genetic algorithm for supervised feature selection. Inf. Sci. 547, 841–859. https://doi.org/10.1016/j.ins.2020.08.083.

Zhu, P., Xu, Q., Hu, Q., Zhang, C., Zhao, H., 2018. Multi-label feature selection with missing labels. Pattern Recogn. 74, 488–502. https://doi.org/10.1016/j.patcog.2017.09.036.