Information Sciences 648 (2023) 119555

Contents lists available at ScienceDirect

Information Sciences

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

FS-MGKC: Feature selection based on structural manifold learning with multi-granularity knowledge coordination

Jie Shi [a,b], Hong Zhao [a,b,*]

[a] School of Computer Science, Minnan Normal University, Zhangzhou, Fujian, 363000, China
[b] Key Laboratory of Data Science and Intelligence Application, Fujian Province University, Zhangzhou, Fujian, 363000, China

ARTICLE INFO

Keywords:
Multi-granularity manifold learning
Feature selection
Knowledge coordination

ABSTRACT

Feature selection uses the hierarchical dependency information provided by multi-granularity knowledge to identify the most relevant features for a given task. Most of these methods manually assign parameters to adjust the restrictive relationships between these dependencies including strong and weak links. However, manual parameter tuning is challenging to accurately determine the influence degree of each parameter to harm performance, because the parameter tuning process involves the interdependence between the multiple hierarchical structures. Along these lines, we propose feature selection based on structural manifold learning with multi-granularity knowledge coordination to balance local structure relationships automatically. First, an affinity matrix based on multi-granularity knowledge is defined as a manifold coordination analysis, which learns all dependencies without destroying the original class space. This analysis effectively trade-offs these dependencies to adjust inter-class alienation and intra-classes inseparable. Afterward, feature correlation and sparsity are leveraged to select features with the most distinguishable and informative features as differentiation analysis. These two analyses jointly selected the feature subsets that are ask-related and informative in the original knowledge structure. Additionally, the proposed method has good universality, which can handle hierarchical classification tasks of the different structure types. Different hierarchical classification results demonstrate the effectiveness of the developed method.

1. Introduction

Increasing numbers of classes and features can inevitably lead to a rapid increase in the size of the classification tasks, which involves tens of thousands or more classes [1]. Coordinating knowledge and information at different class granularities can handle large-scale classification tasks effectively. A commonly used method is the class hierarchical structure, which organizes different categories into different levels, including tree structures and directed acyclic graphs (DAGs), thereby reducing the complexity of classification. The learning tasks guided by hierarchy are called hierarchical classification tasks and are considered a learning process of cognitive structure. Fig. 1 shows two examples to illustrate the hierarchical classification with a tree and a DAG. Although hierarchical classification is widely in many fields [2,3], these classification tasks suffer from the dimensionality curse [4], whereby not all of these high-dimensional features are not relevant for the classification. These tasks contain numerous redundant features,

[*] Corresponding author at: School of Computer Science, Minnan Normal University, Zhangzhou, Fujian, 363000, China.
E-mail address: hongzhaocn@163.com (H. Zhao).

https://doi.org/10.1016/j.ins.2023.119555
Received 5 April 2023; Received in revised form 9 June 2023; Accepted 13 August 2023

Available online 21 August 2023

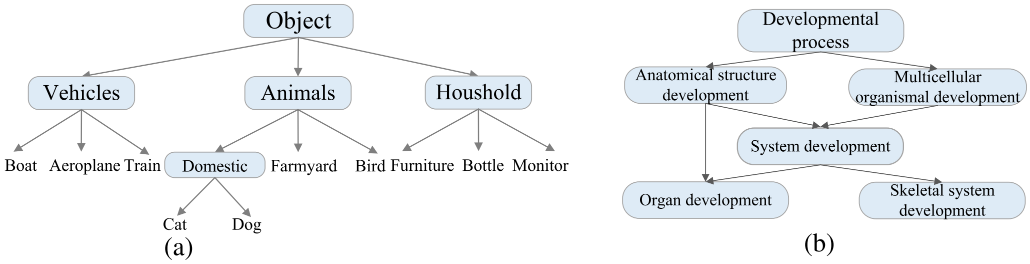


Fig. 1. Two examples of hierarchical classification tasks: (a) the hierarchical classification task with a tree; (b) the hierarchical classification task with a DAG.

aggravating thus the computational burden and degenerating the classification performance. To this end, feature selection can effectively eliminate task-irrelevant features and is widely used to enhance the recognition ability of hierarchical classification tasks [5,6].

The traditional feature selection approach assumes that each class in the classification task is independent. However, these methods are not suitable for performing hierarchical classifications because classes with different granularity have different details and dependencies. Therefore, the feature selection methods considering multi-granularity knowledge are mainly used for large-scale hierarchical classification tasks. These methods select different discriminant features for each class guided by class dependencies, instead of all classes sharing a set of common features. According to the class dependency integrity, the multi-granularity feature selection is divided into feature selections based on single and multi-dependencies. The former only explores the parent-child relationship in the hierarchical classes as a strong link dependency to enhance the similarity constraints [7,8]. The latter fully mines the multi-granularity class knowledge, where the sibling relationship is used as a weak link dependency to supplement the independence between the classes [9,10]. Nonetheless, the above-mentioned methods require manual intervention in the adjustment of class dependency regularization parameters.

However, it is challenging to accurately specify the optimal parameter range by performing manual parameter adjustment. Besides, the parameters cannot be modified even if they are unsuitable for some classification tasks. To solve this problem, manifold learning is used during the feature selection process to reasonably coordinate the relationship between the class dependencies, thus reducing the burden of manual parameter adjustment, such as graph Laplacian [11,12]. The graph Laplacian methods focus on and trade-off different types of structure relationships, which is wildly applied in manifold learning. From the structural dependency domain perspective, the manifold learning-based feature selection is divided into feature selection methods restricted by global and local relationships. As far as the manifold learning-based feature selection procedure is concerned, global relationships [13–15] are considered to capture the global structure relationships by exploring the correlations using different measurement strategies. However, it has been yet proven that local structures play a more crucial role in enhancing the identification ability than local structures. The feature selection guided by local manifold structures [16–18] defines different link dependencies in a unified manifold matrix, which automatically regulates the local discriminative structure relationships conveyed by these dependencies. Based on the above discussion, local manifold learning is embedded into multi-granularity feature selection to automatically intervene in class structure learning and better adjust dependencies.

Under this perspective, this paper proposes feature selection based on structural manifold learning with multi-granularity knowledge coordination (FS-MGKC), which harmonizes the structural dependencies across different granularity classes. First, an affinity matrix is constructed based on the parent-child and sibling relationships of the different granularity classes to capture the dependency relationships between the features, where the matrix serves as coordinated prior knowledge. During each iteration, the coordination of knowledge at different granularity levels is evaluated by analyzing the dependencies captured in the affinity matrix. Based on this analysis, the manifold learning parameters are adjusted to balance the coordination of knowledge until the class structure is stabilized and the coordination of knowledge reaches its optimal state. Furthermore, the feature relationships are considered as feature differentiation analyses to compensate for redundancy caused by manifold coordination. The distance between features is calculated to differentiate the most discriminative features and leverage the sparsity of the $\ell_{2,1}$-norm to differentiate the most informative features. Finally, manifold coordination as a major constraint and feature differentiation analysis as a complementary constraint are simultaneously transformed into mathematical regularizations to guide multi-granularity feature selection.

The contribution is mainly described as follows:

- From a class perspective, structural manifold learning can dynamically tradeoff restriction relationships between the classes to maximize the maintenance of the original semantic information. Semantic information, as a guide to preventing the misclassification of classification models, provides explanations for the subsequent classification process.
- From a feature aspect, the feature selection considering feature sparsity and correlation constraints is regarded as the process that selects the relevant and discards irrelevant features. This leverages a simple and effective feature constraint strategy to effectively eliminate redundancy.
- FS-MGKC was also extended to deal with complex and universal classification tasks with DAGs. From the acquired experience results, it was proven that FS-MGKC is effective for different hierarchical classification tasks.

The remainder of this paper is organized as follows. Section 2 elaborates on the related works. Section 3 describes the FS-MGKC framework and illustrates the optimization objectives. Section 4 explains the experimental setting and discusses the experimental results. Section 5 concludes and proposes further work.

2. Related work

We present the related work, which is divided into three parts as follows: (1) Feature selection based on feature relationships; (2) Feature selection based on separate dependencies; and (3) Feature selection based on unified dependency relationships.

2.1. Feature selection based on feature relationships

Traditional feature selection leverages the different types of feature relationships in the increasingly complex feature space. Some methods use different measure strategies to eliminate the uncorrelated features and select the class-special features with the most discriminative. Jo et al. [19] conducted feature correlation by the improved measurement based on Pearson correlation and R-value for feature redundancy. Similarly, Lim et al. [20] evaluated the feature dependencies with information theory to eliminate redundant features to the maximum extent.

Unlike these methods where an optimization function is not included, the methods based on sparse learning construct a relation regularization term to select the class-common features with the most informative to discriminate all classes. Tibshirani et al. [21] utilized the ℓ_1-norm for sparse regression, which is only designed for binary classification. Various feature selection methods with structural sparsity have been proposed in the literature to solve multi-classification-related problems. Zheng et al. [22] adopted the improved $\ell_{2,0}$-norm to improve the global optimization ability of feature selection. On top of that, to reduce the difficulty of $\ell_{2,0}$-norm parameter adjustment, Pei et al. [23] proposed a robust unsupervised feature selection method based on data relationship learning, where the $\ell_{2,1}$-norm is used to ignore unimportant features. Inspired by this result, we leverage the sparsity of the $\ell_{2,1}$-norm to enforce all tasks share a set of relevant feature subsets.

Nevertheless, all the aforementioned methods only investigate feature information to select a uniform set of discriminant features to recognize all classes, which ignores the multi-granularity knowledge of categories and makes it unable to handle large-scale tasks.

2.2. Feature selection based on separate dependencies

With the rapid development of information technology, the class numbers that need to be recognized in classification tasks have surged, making it impossible to distinguish all classes from a set of features. The feature selection process based on separate class dependencies leverages multi-granularity knowledge and is suitable for performing large-scale classification. Depending on the integrity of the class dependency leveraged, these methods can be catalogized into feature selection guided by single and double dependencies. The feature selection methods based on the single dependency consider one type of granularity-specific dependency. More specifically, Liu et al. [7] introduced a semantic constraint imitating the parent-child relationship to expect a closer distance between the coarse- and fine-grained central classes. Differently, Shi et al. [8] utilized orthogonal constraints to widen the distance between classes, resulting thus in classes with sibling relationships being distributed further in the feature space. These methods based on single dependency only consider the parent-child relationship as a similarity constraint and ignore the sibling relationship as an independence constraint. The feature selection methods based on dual dependency are proposed, considering both dependencies. Zhao et al. [9] presented a feature selection method that comprehensively explores these two class interconnections as similarity and independence constraints for carrying out hierarchical classification. Furthermore, Lin et al. [10] proposed a hierarchical label distribution to represent the correlation between the sibling and parent-child classes, which alleviates the problem of insufficient sampling for minority classes. These methods based on double dependencies comprehensively explore multi-granularity knowledge, where these class dependencies are separately transformed into corresponding constraints.

All the aforementioned methods manually assign the importance of the different dependencies through some experience or domain knowledge, where it is difficult to obtain suitable prior knowledge.

2.3. Feature selection based on unified dependency relationships

Feature selection guided by manifold learning leverages the intrinsic structure and relationships to automatically determine the relevance of features, which offers a valuable alternative to the manual assignment of importance. According to [24], the exploration of manifold learning can be divided into mining global and local structural relationships to enhance discrimination. As far as the global-based methods are concerned, Cai et al. [13] designed manifold-constrained feature selection on feature space, which constructs a graph constraint to explore the intrinsic feature relationship globally. Unlike exploring feature space, Zhang et al. [14] completely captured label structure information by introducing low-dimensional label embedded. Furthermore, Hu et al. [15] considered the connections in both label and feature spaces and constructed dual-graph manifold regularizations to simultaneously mine global affinity. However, the above-mentioned methods explore the global structure information and pay little attention to the local structure information, whereas local relationships provide more specific information.

Unlike the aforementioned global-based methods, Noorie et al. [11] considered local structures constrained by related links to focus on the relationship with maximum information, which ignores the impact of the unrelated links. Therefore, pairwise constraints are introduced as different kinds of link dependencies to automatically distinguish whether each class is related or unrelated

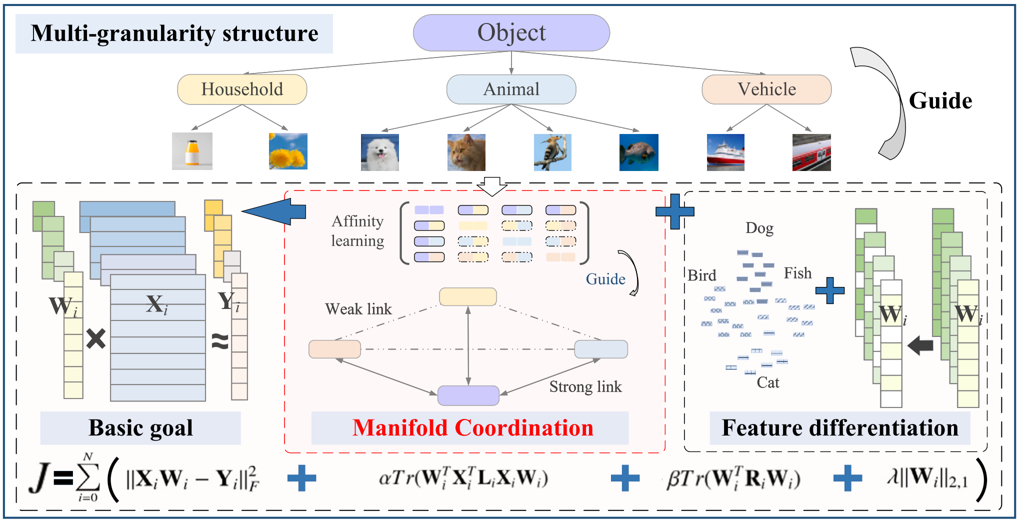

Multi-granularity structure
Object
Household
Animal
Vehicle
Guide
Affinity learning
Guide
Weak link
Strong link
Dog
Bird
Fish
Cat
$\mathbf{W}_i$
$\mathbf{X}_i$
$\mathbf{Y}_i$
$\mathbf{W}_i$
$\mathbf{W}_i$
Basic goal
Manifold Coordination
Feature differentiation
$J=\sum_{i=0}^{N}\left(\|\mathbf{X}_i\mathbf{W}_i-\mathbf{Y}_i\|_F^2+\alpha Tr(\mathbf{W}_i^T\mathbf{X}_i^T\mathbf{L}_i\mathbf{X}_i\mathbf{W}_i)+\beta Tr(\mathbf{W}_i^T\mathbf{R}_i\mathbf{W}_i)+\lambda\|\mathbf{W}_i\|_{2,1}\right)$

Fig. 2. The FS-MGKC framework. Affinity learning establishes a relationship matrix based on the “sibling relationship” and “parent relationship” in a multi-granularity structure to coordinate these relationships. The solid line represents the strong link of the parent-child relationship, while the dashed line represents the weak link of the sibling relationship.

classes [12]. Moreover, manifold learning based on pairwise constraints effectively coordinates local correlations and has been applied to classification tasks. Yin et al. [16] introduced pairwise constraints as prior knowledge to assist multi-view clustering, which utilizes local relationships to effectively regulate inter-class separability and intra-class compactness. Similarly, Rostami et al. [17] adopted pairwise constraints to reduce the ambiguity between the relationships, which actively learns and evaluates different discriminative manifold relationships. Wang et al. [18] designed a multi-manifold regularization term based on pairwise constraints to fully modify the membership relationships of all data objects.

Based on the above-mentioned inspiration, FS-MGKC introduces local manifold learning to capture multi-granularity knowledge, which has the following advantages. (1) The dependency on manual coordination is reduced, as this method automatically adapts to the specific features of each granularity class. (2) The unique relationships and patterns within each granularity level are considered, which allows for a more detailed analysis of the data. (3) The overall performance of the feature selection process is enhanced by capturing the most relevant dependencies and patterns at different granularity levels.

As discussed, although the feature selection methods have been already explored in the literature, the noted methods suffer from the following limitations and challenges.

(1) The feature selection method based on feature relationships ignores the unique requirements and discriminatory information of individual classes. FS-MGKC addresses the internal changes and complexity within different classes by combining multi-granularity knowledge of classes. Thereby, more accurate and fine-grained classification results can be obtained.

(2) The feature selection based on separate dependencies typically requires manual adjustment of the parameters to balance the dependencies between classes, which is both time-consuming and subjective. FS-MGKC automates this process by integrating multi-granularity knowledge coordination, ensuring hence that the class dependencies are considered uniformly without the need for human intervention.

(3) The feature selection method based on unified dependency relationships learns the manifold structure between the features or samples, ignoring thus the dependency relationships of the classes. FS-MGKC unifies the class dependency relationships, which coordinates multi-granularity knowledge and utilizes collective information of all classes. Additionally, FS-MGKC combines information in the feature space to reduce redundancy and enable a comprehensive understanding of the data and the recognition of discriminative features related to multiple classes for performing large-scale classification.

3. Proposed method

First, we review the framework of the feature selection based on structural manifold learning with multi-granularity knowledge coordination (FS-MGKC). Second, we elaborate on the basic problem of multi-granularity feature selection. Thirdly, we introduce the regularization term of FS-MGKC. Finally, we propose the optimization and extension of FS-MGKC.

3.1. The FS-MGKC method framework

The FS-MGKC method is subject to joint constraints from the multi-granularity knowledge and feature relationship perspectives. Fig. 2 illustrates the framework of FS-MGKC.

Details of FS-MGKC are as follows:

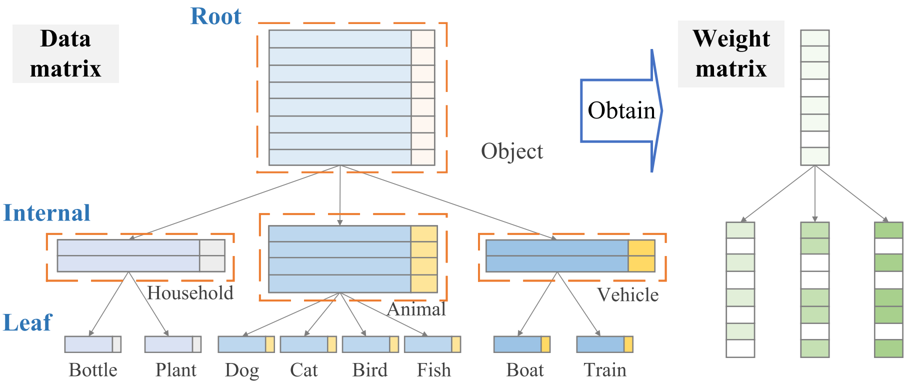


Fig. 3. A classification with a hierarchical tree structure composed of leaf and non-leaf nodes, where non-leaf nodes include a root node and intermediate nodes. The feature weight matrix is obtained by learning the data matrix including the feature (on the left) and label matrix (on the right).

(1) **Manifold coordination regularization**: We focus on the multi-granularity relationships, including class dependencies, to systematically explore strong and weak links among classes. Manifold learning based on multi-granularity knowledge automatically balances these class structural dependencies.

(2) **Feature differentiation regularization**: We leverage the feature correlation and sparse to select features with discriminative and informative information. This process eliminates redundant and irrelevant features produced by structural manifold learning.

3.2. Multi-granularity feature selection

We first introduce the feature selection based on multi-granularity, which is divided into hierarchical structures, task decomposition, and related variables. Then, we mathematically describe the objective of multi-granularity feature selection.

The data of machine learning and data mining often have a hierarchical structure composed of multi-level features, classes, labels, and other information, such as trees and directed acyclic graphs. Fig. 3 depicts a hierarchical tree structure, where these nodes have hierarchical dependencies called parent-child and sibling relationships. For example, nodes “Object” and “Household” are parent-child nodes, and nodes “Household”, “Animal”, and “Vehicle” are sibling nodes.

Handling such large-scale classification tasks requires granularity thinking. A global task is decomposed into related local tasks guided by multi-granularity knowledge, where these local tasks are assigned to non-leaf nodes. To partition local tasks, we comprehensively considered the multi-granularity structure of the dataset and the semantic relationships between classes. A multi-granularity structure provides a hierarchical organization of classes or concepts in a dataset, with each granularity level representing different levels of classification detail or abstraction. WordNet captures semantic relationships between the classes in a dataset [25]. We obtain information about the semantic relationships between classes by mapping class labels or concepts in the dataset to the corresponding WordNet synonym set. Based on this information, we partition each granularity level as an independent local task. For example, we distinguish between “Household” and “Animal” at a coarser granularity level, and then further classified as “Dog” into “Cat”, “Bird”, and “Fish” at a finer granularity level in Fig. 3. Therefore, we have enhanced the construction of the structural dependencies between classes by integrating semantic relationships captured from WordNet. The partitioning of the local tasks simultaneously considers the hierarchical structure and semantic associations between classes, improving thus the overall representativeness and comprehensibility of the dataset.

It is assumed that T_i represents i-th local task and $0 \le i \le (N+1)$, where N is an intermediate node number. From the perspective of hierarchical dependency, the local classification tasks are catalogized into tasks with parent-child and sibling relationships. In local tasks, sibling classes have specific features to distinguish each other, while parent-child classes have shared and similar features. Let $\mathbf{P}$ and $\mathbf{S}$ be the parent-child and sibling set for local tasks, respectively. We focus on local tasks after task decomposition, which reduces the number of classes to distinguish. The identified discriminative features for each local task are also decreased with the reduction of classes, which reduces the classification difficulty of large-scale tasks.

The original data matrix in local task T_i consists of feature matrix $\mathbf{X}_i$ and label matrix $\mathbf{Y}_i$ as shown in Fig. 3. Let m, d, and n_i be the maximum number of classes in all local tasks, the feature number, and the sample number in local task T_i, respectively. Further, we assume that: (1) Feature matrix $\mathbf{X}_i = [\mathbf{x}_1^i; \cdots; \mathbf{x}_j^i; \cdots; \mathbf{x}_{n_i}^i] \in \mathbb{R}^{n_i \times d}$, where $\mathbf{x}_j^i \in \mathbb{R}^{1 \times d}$; (2) Label matrix $\mathbf{Y}_i = [\mathbf{y}_1^i; \cdots; \mathbf{y}_j^i; \cdots; \mathbf{y}_{n_i}^i] \in \mathbb{R}^{n_i \times m}$, where $\mathbf{y}_j^i \in \{0,1\}^{1 \times m}$ corresponds to the labels of a single sample with the element 1 indicates that the sample belongs to a particular class, otherwise it is element 0. In the example shown in Fig. 3, the label matrix $\mathbf{y}_j^i$ of a sample in the local task “Animal” is $[1,0,0,0]$, which means that the sample belongs to the category “Dog”, rather than the classes “Plant”, “Cat”, “Bird”, and’ Fish”. Based on $\mathbf{X}_i$ and $\mathbf{Y}_i$ for local task T_i, we learn feature weight matrix $\mathbf{W}_i \in \mathbb{R}^{d \times m} = [\mathbf{w}_1^i; \cdots; \mathbf{w}_j^i; \cdots; \mathbf{w}_m^i]$. The selection of features strongly depends on the feature weight, which selects features highly related to subsequent tasks and is helpful for classification.

We decompose and simplify the large hierarchical classification task with granularity representation to solve the problem that the feature and feature weight matrixes are not suitable for linear classifiers. Both the feature and feature matrixes in each local task can be linearly combined after the decoupling process. We construct a linear combination $\mathbf{X}_i\mathbf{W}_i$ for local task T_i to represent the

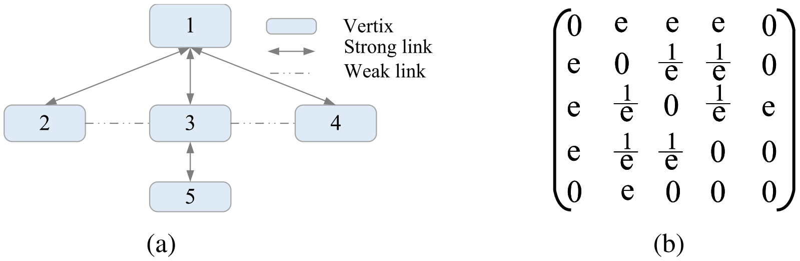


Fig. 4. An example of a hierarchical task with a tree. (a) the corresponding graph with constraints; (b) the corresponding affinity matrix.

predicted labels. The feature selection based on multi-granularity obtains the minimum total error between the predicted and real labels of all local tasks, expressed as

$$L(\mathbf{X}_i, \mathbf{W}_i, \mathbf{Y}_i) = \sum_{i=0}^{N} ||\mathbf{X}_i \mathbf{W}_i - \mathbf{Y}_i||_F^2, \tag{1}$$

where $|| \cdot ||$ is the mathematical expression of the least square.

3.3. Manifold learning based on structural coordination

Manifold learning is widely used in the feature selection procedure to maintain a strong correlation between two classes with strong co-occurrence [26]. Pairwise constraints are also designed to focus on discriminative relationships [27], where strong link constraints make related classes closer and strong link constraints make unrelated classes far away. Inspired by image retrieval, graph Laplacian is widely used in manifold learning for joint consideration of these discriminative dependencies in a unified framework. Therefore, we introduce the graph Laplacian based on dependency constraints, which automatically regulates the local structure dependencies.

We capture the structural manifold by constructing an adjacent graph for class hierarchy, where the node and edge weight represents the class and the affinity between the classes. There are different types of graph construction and weight settings in references [28]. We adopt the scheme that analyzes the relationship between each pair of edges and assigns different weights based on the different relationships. There are usually dependencies between the different classes in hierarchical structures, where the different class dependencies represent link information. It is expected that classes with a parent-child relationship tend to be more connected with each other than classes with a sibling relationship. Therefore, the parent-child and sibling relationships are defined as strong and weak link dependencies, respectively. We define the parent-child and sibling relationships as strong and weak links respectively to better express their relationship at the different granularity layers. Correspondingly, the strong link edge is set at a large weight and the weak link edge is set at a small weight. Therefore, we construct an affinity matrix $\mathbf{M} = [M_{jk}] \in \mathbb{R}^{(N+1)\times(N+1)}$ embedded class dependency, where N denotes the intermediate node number and $0 \le j, k \le N$. Its element M_{jk} is defined as follows:

$$M_{jk} = \begin{cases} e, & \text{if } (j,k) \in \mathbf{P}; \\ \frac{1}{e}, & \text{if } (j,k) \in \mathbf{S}; \\ 0, & \text{otherwise,} \end{cases} \tag{2}$$

where e and $\frac{1}{e}$ are constants. Symbols **P** and **P** represent the parent-child and sibling relationship sets between classes.

Fig. 4(a) illustrates a graph Laplacian corresponding to an example of a classification task with a hierarchical tree in Fig. 1(a). The parent-child relationship represents an intimate relationship related to the adjacent nodes in the hierarchy. Therefore, it is expected that the strong link classes “1” and “2” are more closely related than weak link classes “2” and “3”. The corresponding affinity matrix represents that parent-child classes have greater weight than sibling classes as shown in Fig. 4(b). We obtain that the affinity matrix is symmetrical because parent-child relationships in tree structures are one-to-many.

We construct the discriminative regularization term $R_1(\mathbf{W}_i)$ unifying and regulating different class dependencies formulated as

$$\begin{aligned} R_1(\mathbf{W}_i) &= \frac{1}{2}\sum_{j=1}^{m}\sum_{k=1}^{m}(\mathbf{x}_j^i \mathbf{w}_k^i - \mathbf{x}_k^i \mathbf{w}_k^i)^2 M_{jk} \\ &= \frac{1}{2}\sum_{j=1}^{m}\sum_{k=1}^{m}((\mathbf{w}_k^i)^T \mathbf{x}_j^i + (\mathbf{w}_k^i)^T \mathbf{x}_k^i - ((\mathbf{w}_j^i)^T \mathbf{x}_j^i)((\mathbf{w}_k^i)^T \mathbf{x}_k^i)) M_{jk} \\ &= Tr(\mathbf{W}_i^T \mathbf{X}_i^T \mathbf{F}_i \mathbf{X}_i \mathbf{W}_i) - Tr(\mathbf{W}_i^T \mathbf{X}_i^T \mathbf{M}_i \mathbf{X}_i \mathbf{W}_i), \end{aligned} \tag{3}$$

where $Tr(\cdot)$ is the matrix trace, $\mathbf{F}_i = [f_j^i] \in \mathbb{R}^N$ is a matrix with element $f_j^i = \sum_{j=1}^{N} \mathbf{M}_i$, and $\mathbf{M}_i$ is i-th line of $\mathbf{M}$. We assume that $\mathbf{L}_i = \mathbf{F}_i - \mathbf{M}_i$ and obtain the final manifold coordination constraint $R_1(\mathbf{W}_i)$ formulated as

$$R_1(\mathbf{W}_i) = Tr(\mathbf{W}_i^T \mathbf{X}_i^T \mathbf{L}_i \mathbf{X}_i \mathbf{W}_i). \tag{4}$$

The manifold coordination constraint has the following advantages: (1) It integrates and adaptively trades off all the hierarchical dependencies, and (2) It enhances great compactness within classes and great separability between the classes.

3.4. The FS-MGKC method and optimization algorithm

Under the guidance of the manifold coordination analysis, we obtain the discriminative features related to given tasks. However, these tasks still have redundant features, which reduces the feature diversity. We leverage the different feature strategies to select good features for classification. According to [29], the good features are divided into two categories: (1) Class-common features that have a discriminating effect on all classes; (2) Class-specific features that have identification effects on the different classes for distinguishing each other. From the feature relationship perspective, we explore the feature correlation and sparsity to select these good features.

On the one hand, we calculate the feature correlation to extract class-specific features, eliminating thus redundant features. The relationship between the feature weights in a linear model can reflect the relationship between features and we construct the constraint $R_2(\mathbf{W}_i)$ considering the feature correlation is mathematically expressed as

$$R_2(\mathbf{W}_i) = Tr(\mathbf{W}_i^T \mathbf{R}_i \mathbf{W}_i), \tag{5}$$

where $\mathbf{R}_i = [r_{jk}^i] \in \mathbb{R}^{d\times d}$ is a correlation matrix with element r_{jk}^i represents the distance between feature pairs $\mathbf{x}_j^i$ and $\mathbf{x}_k^i$ defined as

$$r_{jk}^i = \sqrt{\sum_{i=1}^{d} (\mathbf{x}_j^i - \mathbf{x}_k^i)^2}. \tag{6}$$

Eq. (5) effectively reflects the similarity or difference between the feature vectors based on the square root of the differences in each dimension of the feature weight vector, where $\mathbf{R}_i$ quantifies the degree of the difference between feature vectors by measuring the distance between them. Eq. (5) helps to select features that have both independent information and low correlation with each other, thus capturing discriminative information in a given local task.

On the other hand, we adopt the sparse theory to learn class-common features, which enforces that the feature vector irrelevant to the classification is zero [30]. The $\ell_{2,1}$-norm captures features with common informational features for all classes, where the mathematical expression of the $\ell_{2,1}$-norm is

$$R_3(\mathbf{W}_i) = ||\mathbf{W}_i||_{2,1}. \tag{7}$$

Therefore, the final optimization problem based on manifold coordination analysis when further exploring feature relationships is to minimize $J(\mathbf{W}_0, \mathbf{W}_1, \cdots, \mathbf{W}_N)$:

$$J(\mathbf{W}_0, \mathbf{W}_1, \cdots, \mathbf{W}_N) = \sum_{i=0}^{N} (||\mathbf{X}_i \mathbf{W}_i - \mathbf{Y}_i||_F^2 + \alpha Tr(\mathbf{W}_i^T \mathbf{X}_i^T \mathbf{L}_i \mathbf{X}_i \mathbf{W}_i) + \beta Tr(\mathbf{W}_i^T \mathbf{R}_i \mathbf{W}_i) + \lambda ||\mathbf{W}_i||_{2,1}), \tag{8}$$

where the positive parameter α controls the impact of manifold coordination, and the positive parameters β and λ are used to select class-specifical and class-common features. It is noteworthy that the feature weights of all local tasks jointly constitute the feature matrix of hierarchical classification tasks $\mathbf{W} = [\mathbf{W}_0; ...; \mathbf{W}_i; ...; \mathbf{W}_N]$. The top N features after ranking w_j^i in descending order are selected, where value w_j^i represents the j-th feature significance in the i-th local task.

We present an optimization algorithm for Eq. (8). This optimization process is difficult due to the $\ell_{2,1}$-norm non-smoothness. Reference [31] provides the solution for the $\ell_{2,1}$-norm derivation expressed as

$$\frac{\partial(||\mathbf{W}_i||_{2,1})}{\partial \mathbf{W}_i} = 2\mathbf{D}_i \mathbf{W}_i, \tag{9}$$

where $\mathbf{D}_i = [d_{jj}^i] \in \mathbb{R}^{d\times d}$ is a diagonal matrix and $1 \le j \le d$. Its element d_{jj}^i is defined as

$$d_{jj}^i = \frac{1}{2||\mathbf{w}_j^i||_2}, \tag{10}$$

where we set $d_{jj}^i = \epsilon$ when $\mathbf{w}_j^i = 0$. It is worth noting that parameter ϵ is a very small threshold set in advance for calculation accuracy.

We set the derivative of $\mathbf{W}_i$ in Eq. (8) for each local task in hierarchical structures to zero. A derivative expression for $\mathbf{W}_i$ is calculated as

$$\begin{aligned}\frac{\partial J}{\partial \mathbf{W}_i} &= \mathbf{X}_i^T(\mathbf{X}_i \mathbf{W}_i - \mathbf{Y}_i) + \alpha \mathbf{X}_i^T \mathbf{L}_i \mathbf{X}_i \mathbf{W}_i + \beta \mathbf{R}_i \mathbf{W}_i + \lambda \mathbf{D}_i \mathbf{W}_i \\ &= (\mathbf{X}_i^T \mathbf{X}_i + \alpha \mathbf{X}_i^T \mathbf{L}_i \mathbf{X}_i + \beta \mathbf{R}_i + \lambda \mathbf{D}_i)\mathbf{W}_i - \mathbf{X}_i^T \mathbf{Y}_i.\end{aligned} \tag{11}$$

We assume $\mathbf{A}_i = \mathbf{L}_i + \mathbf{L}_i^T$ and $\mathbf{B}_i = \mathbf{R}_i + \mathbf{R}_i^T$, and Eq. (11) is re-formulated as

$$\frac{\partial J}{\partial \mathbf{W}_i} = (\mathbf{X}_i^T\mathbf{X}_i + \alpha\mathbf{X}_i^T\mathbf{A}_i\mathbf{X}_i + \beta\mathbf{B}_i + \lambda\mathbf{D}_i)\mathbf{W}_i - \mathbf{X}_i^T\mathbf{Y}_i. \tag{12}$$

Let the final result of Eq. (12) be zero, we obtain the final expression for $\mathbf{W}_i$ as

$$\mathbf{W}_i = (\mathbf{X}_i^T\mathbf{X}_i + \alpha\mathbf{X}_i^T\mathbf{L}_i\mathbf{X}_i + \beta\mathbf{R}_i + \lambda\mathbf{D}_i)^{-1}(\mathbf{X}_i^T\mathbf{Y}_i). \tag{13}$$

We briefly describe and analyze FS-MGKC in Algorithm 1, where the symbol $O(\cdot)$ presents the time complexity of the operation.

Algorithm 1 The algorithm of FS-MGKC.

Input: (1) Feature matrix $\mathbf{X}_i \in \mathbb{R}^{n_i\times d}$ and label matrix $\mathbf{Y}_i \in \mathbb{R}^{n_i\times m}$; (2) Regularization parameters: λ, α, and β; (3) Parameter $i \in \{0, 1, \cdots, N\}$, where N is the intermediate node number; (4) The maximal iteration number T;

Output: Weight matrix $\mathbf{W} \in \mathbb{R}^{n\times d}$;

```
1: for $i = 0 : N$ do
2:     initialize $\mathbf{W}_i \in \mathbb{R}^{d\times m}$ randomly;
3:     Compute the feature relation matrix $\mathbf{R}_i$ according to Eq. (6);
4: end for
5: Compute affinity matrix $\mathbf{M}$ according to Eq. (2);
6: $\mathbf{W} = [\mathbf{W}_0, \mathbf{W}_1, \cdots, \mathbf{W}_N]$;
7: $t = 0$;
8: while $t < T$ do
9:     for $i = 0 : N$ do
10:        Compute the diagonal matrix $\mathbf{D}_i^{(t)}$ according to Eq. (10);
11:        Compute the diagonal matrix $\mathbf{F}_i^{(t)}$ according to $j$-th element $F_j^i = \sum_{j=1}^N \mathbf{M}_i$;
12:        Compute $\mathbf{L}_i^{(t)} = \mathbf{F}_i^{(t)} - \mathbf{M}_i^{(t)}$;
13:        Update $\mathbf{W}_i$: $\mathbf{W}_i^{(t+1)} = (\mathbf{X}_i^T\mathbf{X}_i + \lambda\mathbf{D}_i^{(t)} + \alpha\mathbf{X}_i^T\mathbf{L}_i^{(t)}\mathbf{X}_i + \beta\mathbf{R}_i)^{-1}(\mathbf{X}_i^T\mathbf{Y}_i)$;
14:    end for
15:    Update $\mathbf{W}^{(t+1)} = [\mathbf{W}_0^{(t+1)}, \mathbf{W}_1^{(t+1)}, \cdots, \mathbf{W}_N^{(t+1)}]$;
16:    $t = t + 1$;
17: end while
18: return $\mathbf{W}$;
```

Besides, m, d, and n_i can be defined as the maximum class number, the feature number, and the sample number in i-th local task, respectively. The FS-MGKC time complexity is mainly subject to the $\mathbf{W}_i$ update process for the non-leaf tasks in line 13. We construct a feature relation matrix $\mathbf{R}_i$ for i-th node and an affinity matrix $\mathbf{M}$ in advance, where $\mathbf{M}_i$ represents the i-th line of $\mathbf{M}$. The diagonal matrix $\mathbf{D}_i^{(t)}$ for i-th node is computed. The calculations for $\mathbf{X}_i^T\mathbf{X}_i$, $\mathbf{X}_i^T\mathbf{Y}_i$, and $\mathbf{X}_i^T\mathbf{L}_i\mathbf{X}_i$ are the main time consumptions, which require $O(d^2n_i)$, $O(dn_im)$ and $O(d^2n_i)$. Therefore, the total update for all non-leaf tasks requires $O(d^2n)$, $O(dnm)$, and $O(d^2n)$. The time complexities for calculating matrixes $\mathbf{R}_i$, $\mathbf{M}$, and $\mathbf{D}_i$ are far less than these main time complexities, so they can be ignored. Therefore, the final FS-MGKC time complexity is $O(T(d^2n + dnm))$, and T refers to the total iteration number.

The FS-MGKC method has the following advantages: (1) The local manifold structure based on class dependencies is modeled by a graph Laplacian. This process adaptively adjusts these class dependencies to achieve balance based on embedding class structure dependencies; (2) Feature correlation is considered to eliminate the features with duplicate information.

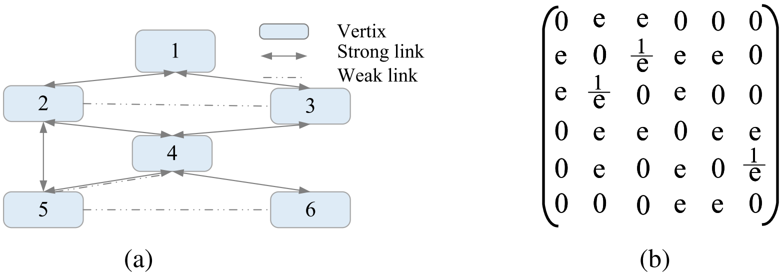
3.5. Method expansion and discussion

We expand FS-MGKC to directly deal with hierarchical classification tasks with DAGs. These hierarchical DAG structures, similar to hierarchical tree structures, have parent-child and sibling dependencies. Unlike the hierarchical tree, these dependencies of DAGs have the following properties: (1) The parent-child relationship of a DAG is a type of many-many relationship; (2) The sibling and parent-sibling relationships may contradict each other. We use an example to clearly illustrate these cases, where Fig. 5(a) depicts a graph Laplacian corresponding to a classification task with a DAG in Fig. 1(b). For example, class “4” has two parent classes “2” and “3”. Classes “4” and “5” have both sibling and parent-child relationships. The parent-child relationship has the “strong link” ability and the sibling relationship has the “weak link” ability. We believe that the parent-child relationship takes precedence over the sibling relationship if these relationships exist. Fig. 5(b) shows an affinity matrix corresponding to the hierarchical task with a DAG, which is asymmetric.

The final optimization for DAGs is to minimize $J(\mathbf{W}_0, \mathbf{W}_i, \cdots, \mathbf{W}_N)$ in Eq. (8), where the definition of manifold preserving matrix is different from that of trees.

The FS-MGKC extension has the following advantages: (1) DAGs reflect more comprehensive and complex information, such as class relationships spanning multiple layers, while trees only reflect local class relationships. Therefore, DAGs provide more comprehensive structural information. (2) DAGs provide more comprehensive and rich feature information since each node considers the information of all connected nodes.

The FS-MGKC method is designed to handle hierarchical classification tasks, but there are some difficulties in handling large-scale graph structure classification tasks. There are several reasons why FS-MGKC is difficult to perform large-scale DAG classification: (1) High computational complexity: The computational complexity of affinity matrices used to construct graph-structured datasets increases with the number of classes and instances. (2) Complexity of DAG structure: DAG has complex and diverse structures



$$\begin{pmatrix} 0 & e & e & 0 & 0 & 0 \\ e & 0 & \frac{1}{e} & e & e & 0 \\ e & \frac{1}{e} & 0 & e & 0 & 0 \\ 0 & e & e & 0 & e & e \\ 0 & e & 0 & e & 0 & \frac{1}{e} \\ 0 & 0 & 0 & e & e & 0 \end{pmatrix}$$

Fig. 5. An example of a hierarchical task with a DAG structure. (a) the interpretive graph with the corresponding constraints; (b) the corresponding affinity matrix.

Table 1
Descriptions of experimental datasets with hierarchical structures.

Dataset	Type	Feature	Training	Test	Class
DD	Tree	473	3,020	605	27
F194	Tree	473	7,105	1,420	194
Bridge	Tree	11	84	24	6
CLEF	Tree	80	8368	939	63
AWA	Tree	252	6,405	3,202	10
VOC	Tree	4,096	3,437	3,539	20
Cifar	Tree	4,096	50,000	10,000	100
ILSVRC	Tree	4,096	12,346	11,845	57
SUN	Tree	4,096	45,109	22,556	324
Eisen	DAG	80	1048	820	1707
Derisi	DAG	64	1598	1255	2037
Cellcycle	DAG	78	1612	1274	2041

in large-scale classification tasks. The FS-MGKC method relies on accurately capturing the hierarchical relationships between the classes within the DAG. However, effectively modeling and representing these relationships becomes increasingly challenging as DAG becomes increasingly complex, affecting thus FS-MGKC's performance.

4. Experimental setup and results analysis

We illustrate the experimental setup, which is described by dataset descriptions, experimental comparison methods, and experimental settings. We compare performance from effectiveness and efficiency. Our experiments include (1) Effectiveness comparison experiments with other methods; (2) Efficiency experiments compared with other methods; (3) Influence of the selected feature number for FS-MGKC; (4) Influence of different regularization constraints for FS-MGKC; (5) Extended experiment of FS-MGKC on DAG datasets; (6) Parameter sensitivity analysis; and (7) Convergence analysis.

The experiment configuration is as follows: a PC with Intel Core i7-3770, 3.40 GHz CPU, 32.0 GB memory, and 64-bit Windows 10 operating system.

4.1. Experimental dataset descriptions and comparison methods

First, we introduce and describe the experimental datasets with different hierarchical structures. According to different hierarchies, these datasets are divided into datasets with tree and DAG structures. These datasets with tree structures include *CLEF* [32], *AWA* [33], *VOC* [34], *Cifar* [35], *ILSVRC* [36], *SUN* [37], *DD* [38], and *F194* [39], where they are introduced from different fields: protein and image. These datasets with DAG structures include three gene datasets: *Eisen* [40], *Derisi* [40], and *Cellcycle* [40], which are used to construct the extended experiment of FS-MGKC. We describe these datasets, and the details are listed in Table 1.

Second, we compare FS-MGKC with some hierarchical methods. The comparison methods can be divided into two categories: (1) Hierarchical feature selection methods based on feature relationships *HFisher* [41], *HFSNM* [42], and *HMRMR* [43]. (2) Hierarchical feature selection methods based on class dependency relationships *HiFSRR* [9], *HFSDK* [7], *HFSCN* [44], *HFSGR* [45], and *LCCSHFS* [46], where *LCCSHFS* is an extended method for DAG. Detailed descriptions of these methods are as follows:
(1) *HFisher* is a label distance-based feature selection method guided by *Fisher*.
(2) *HFSNM* jointly sparse the loss and regularization terms in the hierarchical classification to select the most representative features.
(3) *HMRMR* uses information theory to embed feature relationships to assist feature selection based on multi-granularity.
(4) *HiFSRR* is inclined to fully exploit the class dependency relationships for the discriminative feature of multi-granularity classification.
(5) *HFSDK* is inspired by the idea of the parent-child relationship to reduce the negative impact of the outliers in multi-granularity classification.

Table 2
The experimental results on different tree datasets with different feature selection methods in terms of F_{LCA} (↑).

Method	*DD*	*F194*	*Bridge*	*CLEF*	*AWA*	*ILSVRC*	*VOC*	*Cifar*	*SUN*
HimRMR	82.01	62.81	69.67	68.07	46.44	91.87	74.99	74.13	75.09
HiFSNM	80.94	56.94	70.17	65.88	48.45	90.34	74.38	–	–
HiFisher	74.70	59.32	68.00	67.76	48.15	91.72	74.92	73.83	74.64
HFSCN	81.72	60.20	67.00	66.47	48.66	85.50	74.54	–	–
HFSDK	81.79	61.49	66.17	66.51	48.48	85.81	74.69	–	–
HFSGR	81.86	62.17	66.00	66.64	48.14	86.02	74.74	–	–
LCCSHFS	81.73	63.16	67.00	67.34	48.38	90.92	74.85	73.93	76.01
HiFSRR	80.80	63.40	67.83	67.15	48.71	91.00	74.92	73.98	**76.31**
FS-MGKC	**82.97**	**63.94**	**71.33**	**68.21**	**48.80**	**91.98**	**75.17**	**74.23**	76.12

(6) *HFSCN* resist noise data using a capped ℓ_2-norm to select discriminative features in multi-granularity tasks.
(7) *HFSGR* considers the directionality of class dependency, which learns the bidirectional inter-class dependencies using the subtree graph to provide additional classification information.
(8) *LCCSHFS* is inclined to analyze the label relevance and embeds these analyses in different hierarchical structures, including trees and DAGs.

Unlike these methods, FS-MGKC explores the class dependency-based manifold learning to better understand the correlation between classes in data, and combine it with the relationship between features. Moreover, FS-MGKC extends its applicability to DAG tasks, making it suitable for a broader range of real-world problems.

4.2. Experimental settings

We present the experimental settings described from two aspects: the settings for evaluation indicators and parameters.

Evaluation indicators assess the effectiveness and efficiency. On the one hand, the Hier-ACC [45], Hier-F_1 [47], F_{LCA} [48], and TIE [49] indicators describe the effectiveness. Hier-ACC is a weighted accuracy indicator for performing hierarchical classification. Hier-F_1 considers both predict and real classes of the ancestors and descendants. For observation convenience, we expand the Hier-F_1 value by 100-fold. F_{LCA} uses graph theory to represent the distance between the predicted and real labels. TIE represents the error edge between the actual and the predicted labels in hierarchical class structures. On the other hand, the feature selection efficiency is assessed by the running time.

Experimental parameters are set in advance for discriminative feature selection from the original data. The experimental parameter descriptions can be mainly divided into two aspects. First, we introduce the parameter settings. (1) The comparison method parameters follow the original paper setting. (2) The FS-MGKC parameters: α and λ are set to 10, and β is set to 1. Parameter α represents the adjustment strength of the manifold coordination term. A too large α value causes the feature to pay too much attention to the task and ignore the generation of duplicate information. On the contrary, a too-small α value ignores the guidance of multi-granularity knowledge and degenerates into traditional feature selection. The parameters β and λ represent the strength of the feature discrimination regularization term. The β and λ values are too large to delete many classification-related features and too small to reduce redundant information.

Second, we introduce the parameter settings of the classifier in the testing process. A support vector machine (SVM) classifier evaluates classification performance, where the selected features are conducted by 10-fold cross-validation.

4.3. Effectiveness comparison experiments with other methods

We evaluate the quality of the features selected by FS-MGKC and other methods in terms of the, F_{LCA}, TIE, and Hier-F_1 indicators. Symbol "–" indicates insufficient memory capacity to deal with this classification.

First, we select 10%, 20%, 30%, 40%, and 50% of the total feature number using different methods and adopt Hier-ACC to evaluate feature qualities shown in Fig. 6.

A higher Hier-ACC value indicates better performance. From Fig. 6, we obtain that FS-MGKC exhibits good performance on almost all datasets. We discuss the following detailed conclusions:

(1) The FS-MGKC method exhibits excellent performance in different types of classification tasks, including the protein and image datasets.

(2) The coordination of FS-MGKC on multi-granularity knowledge gradually emerges performance advantages in large-scale datasets with the increase in the selected features. FS-MGKC explores and harmonizes class structure information of *Cifar* and *SUN* to select identification features. FS-MGKC has also a remarkable performance on simple datasets with few features. For example, the FS-MGKC performance is significantly better than other methods when 10% of all features are selected on the *F194* dataset.

As can be seen from Fig. 6, selecting 10% of all features can represent the performance difference between these methods. Therefore, each method selects 10% of the features in the subsequent experiments.

Then, we assess the discriminant ability of these methods in terms of the F_{LCA} and TIE indicators. Table 2 lists the F_{LCA} indicator with different feature selection methods on tree datasets, where symbol "↑" represents that the higher the better.

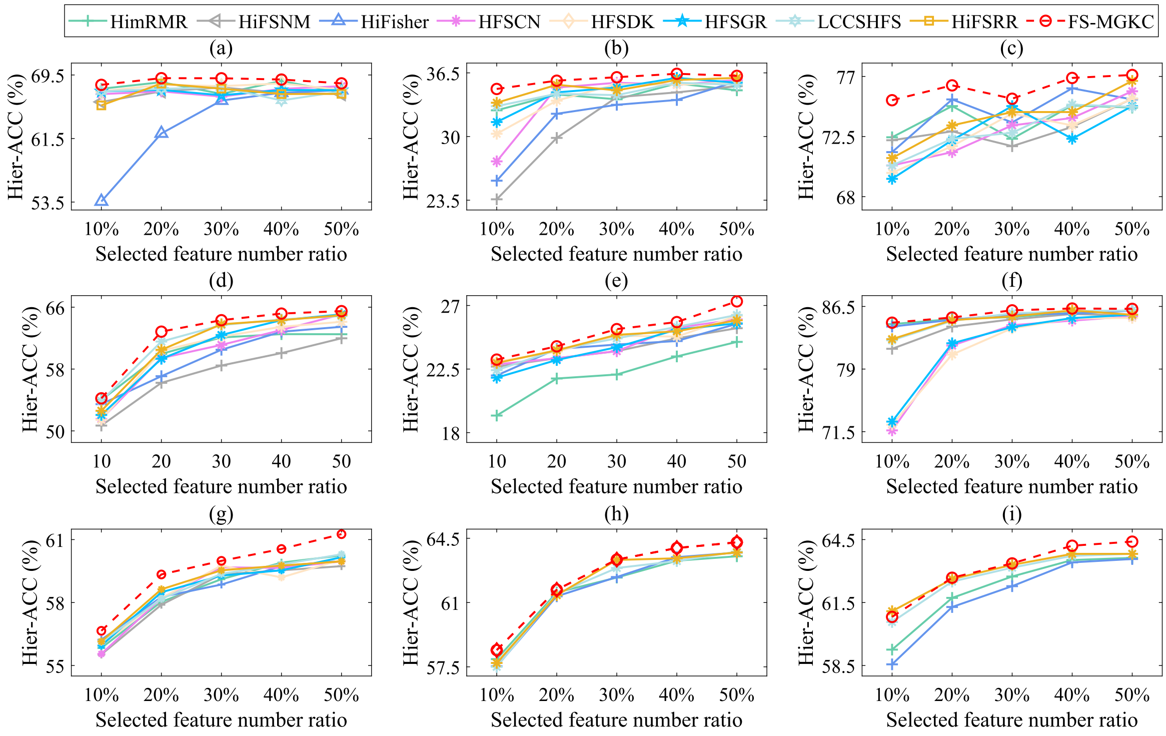


Fig. 6. The experimental results with different feature selection methods in terms of Hier-ACC. (a) *DD*; (b) *F194*; (c) *Bridge*; (d) *CLEF*; (e) *AWA*; (f) *ILSVRC*; (g) *VOC*; (h) *Cifar*; and (i) *SUN*.

Table 3
The experimental results on different tree datasets with different feature selection methods in terms of TIE (↓).

Method	*DD*	*F194*	*Bridge*	*CLEF*	*AWA*	*ILSVRC*	*VOC*	*Cifar*	*SUN*
HimRMR	52.6	251.4	3.5	209.4	1145	419.6	439.6	1421	3196
HiFSNM	56.4	299.8	3.4	222.3	1112	497.4	451.8	–	–
HiFisher	71.4	270.0	3.7	210.5	1115	425.8	439.1	1442	3258
HFSCN	53.2	267.2	3.7	220.0	1106	721.6	445.4	–	–
HFSDK	53.4	260.2	3.7	218.4	1115	710.4	444.0	–	–
HFSGR	52.6	255.4	3.7	218.6	1114	696.4	442.7	–	–
LCCSHFS	53.8	245.0	3.7	212.9	1112	462.4	436.0	1429	3053
HiFSRR	56.4	245.6	3.7	215.7	1105	457.6	439.7	1431	**3018**
FS-MGKC	**50.4**	**244.4**	**3.1**	**209.4**	**1104**	**411.4**	**439.0**	**1415**	3047

Table 3 lists the TIE indicator on tree datasets with different feature selection methods, where "↓" indicates that the lower the better.

The F_{LCA} and TIE indicators are unique indicators for evaluating hierarchical methods. As can be observed from Tables 2 and 3, the following conclusions are obtained:

(1) The results display that the discrimination ability of FS-MGKC exceeds other classical hierarchical methods without class relationships on all datasets. This result shows that using different levels and granularity of knowledge can more accurately capture the features in the data, and improve the algorithm performance in many scenarios.

(2) The FS-MGKC performance is better than the five advanced methods with all methods investigating class structure dependencies. This result demonstrates that FS-MGKC better coordinates multi-granularity knowledge to fully mine class dependencies on most datasets.

Finally, we compare the Hier-F_1 indicators for different methods. Table 4 lists the Hier-F_1 indicators of different methods, where a higher Hier-F_1 value indicates better performance.

From Table 4, the following conclusions can be summarized:

(1) The selection ability for distinguishing features of FS-MGKC is excellent on most datasets from the Hier-F_1 perspective. The Hier-F_1 indicator is a unique indicator for evaluating hierarchical methods. Therefore, FS-MGKC can better understand and process complex data and knowledge and improve the accuracy of classification tasks.

(2) The FS-MGKC method has no obvious performance advantage on the *Cifar* and *SUN* datasets. Due to the insufficient weight given to manifold coordination, FS-MGKC is challenging to reconcile the excessive class dependencies in large datasets.

Table 4
The experimental results on different tree datasets with different feature selection methods in terms of Hier-F_1 (↑).

Method	*DD*	*F194*	*Bridge*	*CLEF*	*AWA*	*ILSVRC*	*VOC*	*Cifar*	*SUN*	Rank
HimRMR	85.51	70.49	72.44	70.91	55.31	93.07	78.31	76.32	82.29	3.8
HiFSNM	84.46	64.81	72.22	68.90	56.58	94.75	77.81	–	–	7.0
HiFisher	80.34	68.31	71.33	70.73	56.46	95.51	78.41	75.97	81.95	5.4
HFSCN	85.36	68.64	70.33	69.50	56.82	92.38	78.06	–	–	6.5
HFSDK	85.29	69.46	69.78	69.58	56.46	92.50	78.07	–	–	6.9
HFSGR	85.22	70.02	69.33	69.62	56.51	92.65	78.25	–	–	6.6
LCCSHFS	85.18	70.92	70.33	70.46	56.61	95.12	78.31	76.18	83.08	4.1
HiFSRR	84.49	71.17	70.89	70.14	56.85	95.17	78.58	76.15	**83.27**	3.6
FS-MGKC	**86.11**	**71.31**	**75.22**	**70.94**	**56.91**	**95.66**	**78.63**	**76.64**	83.11	1.1

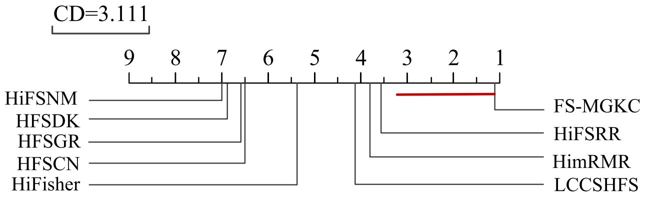


Fig. 7. Average ranks of all methods with the critical distance (CD) for Hier-F_1.

Table 5
Running time (s) of different feature selection methods, where the best results are highlighted in bold.

Method	*DD*	*F194*	*Bridge*	*CLEF*	*AWA*	*ILSVRC*	*VOC*	*Cifar*	*SUN*	Rank
HimRMR	22.3	51.9	0.034	3.01	18.3	11236	2897	25001	36665	6.7
HiFisher	8.56	55.2	0.043	**0.09**	55.4	452.4	**1.01**	–	–	5.6
HiFSNM	**0.16**	**0.89**	0.023	85.5	**0.14**	1483	17.3	**8.908**	**19.91**	2.8
HFSCN	1.44	9.79	0.045	5.43	6.47	196.9	32.1	–	–	5.9
HFSDK	11.8	82.1	0.043	35.5	56.8	1887	193	–	–	7.6
HFSGR	0.48	1.19	0.066	0.14	0.26	**4.111**	65.2	–	–	4.0
LCCSHFS	1.73	4.89	0.115	0.85	1.45	149.9	914	5016	5024	5.4
HiFSRR	0.93	3.75	0.123	0.38	0.42	83.18	90.4	2319	871.7	4.2
FS-MGKC	0.62	2.82	**0.016**	0.30	0.74	129.5	46.3	402.8	347.4	2.8

Furthermore, we adopt the statistical methods to compare the Hier-F_1 indicator differences between FS-MGKC and other methods. First, we leverage the Friedman test to determine the existence of differences. We define the method numbers $K = 9$ and experiment datasets $N = 9$. Next, we compute the average rank R in terms of the Hier-F_1 indicator for the different feature selection methods. We assume that the null hypothesis represents all feature selection method ranks are equivalent. Under this premise, Friedman statistics is represented as $F_f = \frac{(N-1)\chi_F^2}{N(K-1)-\chi_F^2}$, where $\chi_F^2 = \frac{12N}{K(K+1)}(\sum_{i=1}^{N} R_i^2 - \frac{K(K+1)^2}{4})$. We assume freedom degrees of F_f are $(K-1)$ and $(K-1)(N-1)$ and it is subject to the F distribution. We obtain that $F_f = 7.574$ and $F((9-1),(9-1)(8-1)) = 2.087$ when the significance level α is 0.05. Therefore, there are obvious differences among these nine methods, which represents the null hypothesis is false.

Then, the Bonferroni Dunn test [50] is used to systematically explore the different degrees. We define critical distance $CD_\alpha = q_\alpha\sqrt{\frac{K(K+1)}{6N}}$ and calculate $CD = 3.111$ when $\alpha = 0.05$ and $q_\alpha = 2.724$. The statistical results of different methods with the Bonferroni Dunn test are shown in Fig. 7. From our analysis, we obtain that the critical distance is less than the distance between the average ranks, indicating that these methods have significant differences.

As can be ascertained from Fig. 7, it can be intuitively observed that: (1) There are obvious differences between FS-MGKC and other hierarchical methods. (2) The statistical results on multiple datasets show that the FS-MGKC performance is significantly better than the other methods.

4.4. Efficiency experiments compared with other methods

We compare the efficiency discrepancy between FS-MGKC and other methods. The running time results on all experimental datasets are listed in Table 5. It can be concluded that FS-MGKC has excellent efficiency in dynamically balancing multi-granularity knowledge on large datasets, while FS-MGKC efficiency advantage is not prompt on small datasets.

We further analyze the differences in efficiency between the FS-MGKC and other hierarchical methods from the statistical aspect. Table 5 lists the average ranking of different feature selection methods in running time. We define the significance level $\alpha = 0.05$ and calculate $F_f = 4.73$ under the null hypothesis, indicating that the null hypothesis is not satisfied. Therefore, all methods are

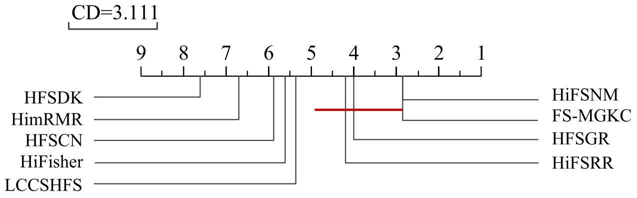


Fig. 8. Average rank results of all methods with the critical distance (CD) for running time.

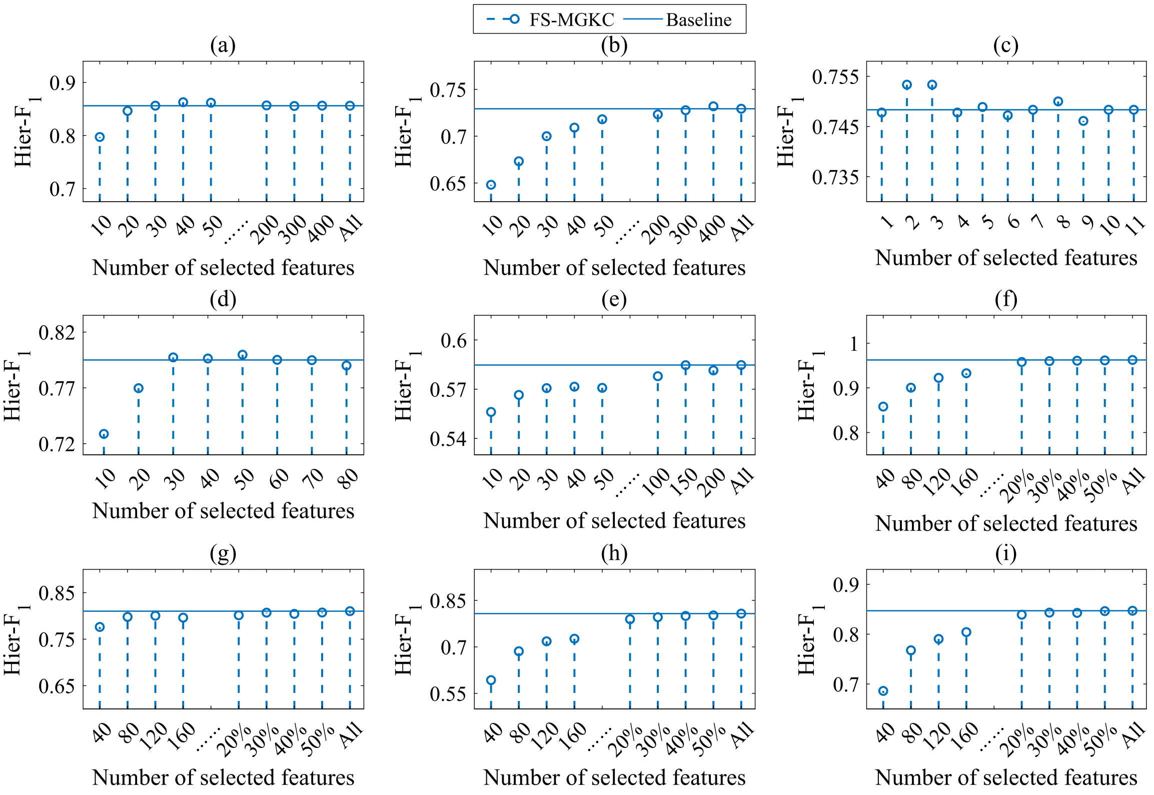


Fig. 9. The experimental results for features selected by FS-MGKC in terms of Hier-F_1. (a) *DD*; (b) *F194*; (c) *Bridge*; (d) *CLEF*; (e) *AWA*; (f) *ILSVRC*; (g) *VOC*; (h) *Cifar*; and (i) *SUN*.

unequal. We explore the efficiency differences with the Bonferroni Dunn test [50]. We calculate the critical distance $CD_\alpha = 3.111$ when $\alpha = 0.05$. Fig. 8 shows the statistical results of the running time with the Bonferroni Dunn test.

From Fig. 8, we obtain: (1) FS-MGKC is superior to HFSDK, HimMRR, HiFisher, and HFSCN from the efficiency perspective; (2) there is no concrete piece of evidence of significant differences between FS-MGKC and HiFSNM, HFSGR, and HiFSRR in terms of efficiency.

4.5. *Influence of the selected feature number*

We discuss the effect of the discriminant feature number on FS-MGKC performance. The FS-MGKC method selects different numbers of features as different comparative experiments and all features as baseline experiments. A 10-fold cross-validation SVM classifier tests and evaluates these features in terms of the Hier-F_1 indicator. Fig. 9 shows Hier-F_1 with the different number of features selected by FS-MGKC.

From Fig. 9, the following conclusions can be drawn:

(1) The increase in the selected feature number results in the FS-MGKC performance is closer to the baseline experimental performance. This reflects that FS-MGKC can select discriminative features.

(2) The FS-MGKC ability to select discriminative features is non-prominent when the selected feature number is few. For example, FS-MGKC selects 60% of the features to achieve good classification on the *AWA* dataset.

(3) The FS-MGKC classification performance is well reflected on large-scale datasets, where about 10% to 20% of the selected features achieve the baseline experiment effect. A notable example is the *SUN* dataset, where about 10% of the features realize the classification accuracy of all features.

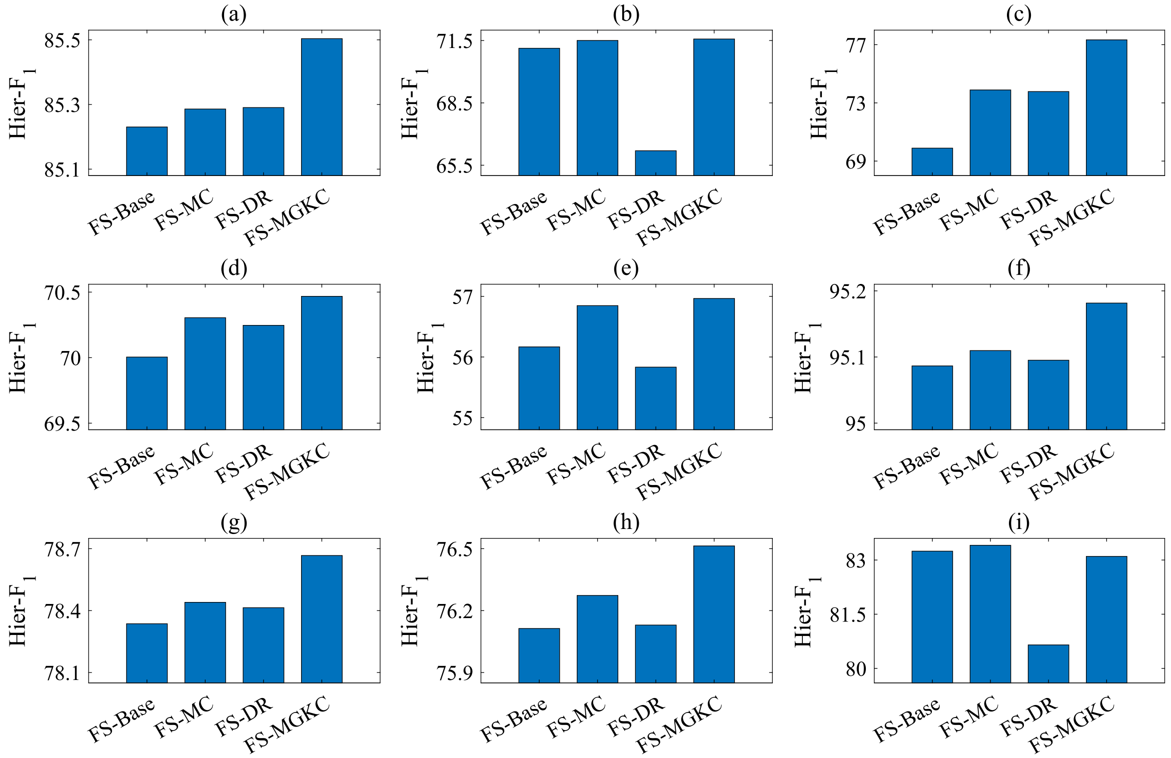


Fig. 10. The experimental results for different FS-MGKC regularizations in terms of Hier-F_1. (a) *DD*; (b) *F194*; (c) *Bridge*; (d) *CLEF*; (e) *AWA*; (f) *ILSVRC*; (g) *VOC*; (h) *Cifar*; and (i) *SUN*.

(4) The FS-MGKC method is effective for datasets in different fields. For example, selecting 20 features on the protein *DD* dataset and 80 features on the image *VOC* dataset completes satisfactory classification performance.

4.6. Influence of different regularization constraints

We discuss the influence of each regularization term in FS-MGKC and use the following method to intuitively compare the impact of each regularization term. (1) FS-Base combines the loss function and $\ell_{2,1}$-norm regularization term as the baseline experience. (2) FS-MC adds the manifold coordination regularization based on FS-Base to constrain class dependency. (3) FS-DR considers the differentiation regularization for specific-class features to reduce redundancy. (4) FS-MGKC simultaneously utilizes these two regularization terms based on the baseline experiment. For the experiment fairness, the above-mentioned four methods select 10% of the features on each dataset and evaluate these features with Hier-F_1. Fig. 10 illustrates the experimental results for the different FS-MGKC regularizations in terms of Hier-F_1.

As can be seen in Fig. 10, the following conclusions can be observed and summarized:

(1) The experimental results indicate that the manifold coordination regularization positively improves the performance to effectively trade off the different class dependencies. The performance of HFS-MR is superior to that of HFS-Base on all datasets.

(2) The performance of HFS-DR is superior to that of HFS-Base on most datasets, except *F194*, *AWA*, and *SUN*. The experimental results demonstrate that the feature discrimination regularization term largely regulates the relationship between the features. Feature spaces of *F194* and *AWA* are relatively simple, whereas an excessive constraint on the feature differentiation regularization leads to poor performance. The regulating ability of the feature discrimination regularization term on *SUN* is insufficient.

(3) The results indicate that the combination of manifold coordination and feature discriminant regularization can effectively select the optimal feature subset. The performance of HFS-MIMR exceeds that of the other three methods on all datasets. The manifold coordination regularization identifies the highly task-related feature subset, and the feature constraint further removes the information-redundant features.

4.7. Extended experiment of DAG structures

We verify the FS-MGKC effectiveness expansion by using four gene datasets with DAGs. We compared the ability of the FS-MGKC and LCCSHFS methods to deal with different DAGs datasets. We vary the ratio of the selected feature number and record different Hier-F_1 values in 10%, 20%, 30%, 40%, 50%. Table 6 lists the experimental results on DAG datasets using the FS-MGKC and LCCSHFS methods in terms of the Hier-F_1 indicator.

Table 6
The experimental results on DAG datasets using different methods in terms of Hier-F_1.

Dataset	Method	Selected feature number				
		10%	20%	30%	40%	50%
Eisen	FS-MGKC	35.18	36.71	37.44	37.96	37.48
	LCCSHFS	34.59	36.40	37.15	37.52	37.71
Cellcycles	FS-MGKC	26.37	27.88	28.74	30.42	31.87
	LCCSHFS	26.24	27.74	28.16	30.00	31.64
Derisi	FS-MGKC	26.71	26.87	26.86	26.63	26.95
	LCCSHFS	26.29	26.04	26.86	27.38	26.72

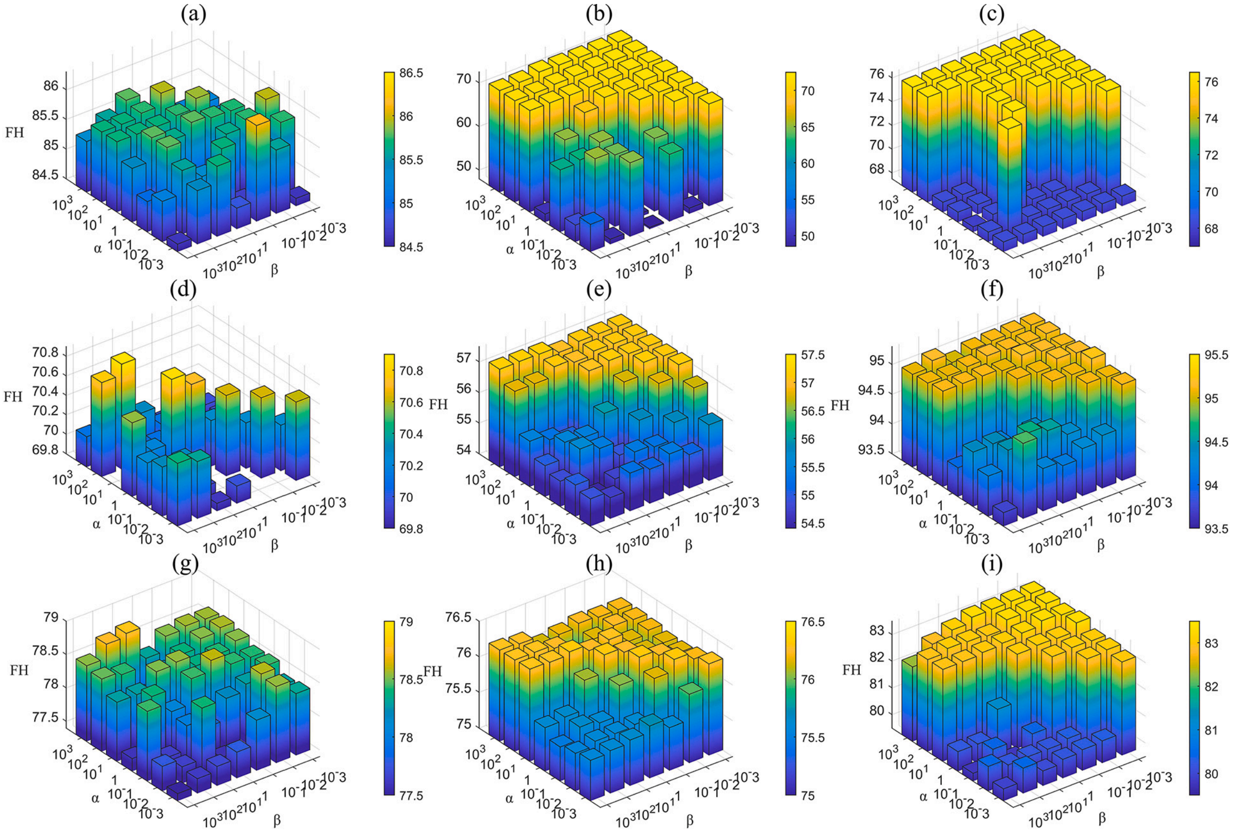


Fig. 11. Parameter sensitivity analysis. (a) *DD*; (b) *F194*; (c) *Bridge*; (d) *CLEF*; (e) *AWA*; (f) *ILSVRC*; (g) *VOC*; (h) *Cifar*; and (i) *SUN*.

From Table 6, we obtain the followings: (1) The FS-MGKC method coordinates the multi-granularity knowledge in DAG to handle more complex structure classification tasks. (2) The FS-MGKC method exceeds the LCCSHFS method in the ability to select discriminating features on the toy DAG datasets.

4.8. *Parameter sensitivity analysis*

We discuss the sensitivity of the FS-MGKC parameters on different datasets, where λ controls feature sparsity, β adjusts the feature correlation, and α regulates manifold structure. We assume that feature sparsity is a basic factor and λ is limited to 10. Parameters α and β are varied over the set $\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10^1, 10^2, 10^3\}$. Fig. 11 reflects the parameter variety trend of FS-MGKC performance.

From Fig. 11, the following can be observed and summarized:

(1) Parameter α controls the coordination degree of the structural manifold regularization. FS-MGKC has poor performance when $\alpha < 1$, a small α value results in an insufficient manifold coordination degree, and destroys local structural information to endanger performance. FS-MGKC selects the distinguishing features more related to the given task when $\alpha > 1$, where its performance is relatively stable with the growth of the α value. The above-mentioned situation reveals that larger values have a more positive effect on the manifold coordination of local structures.

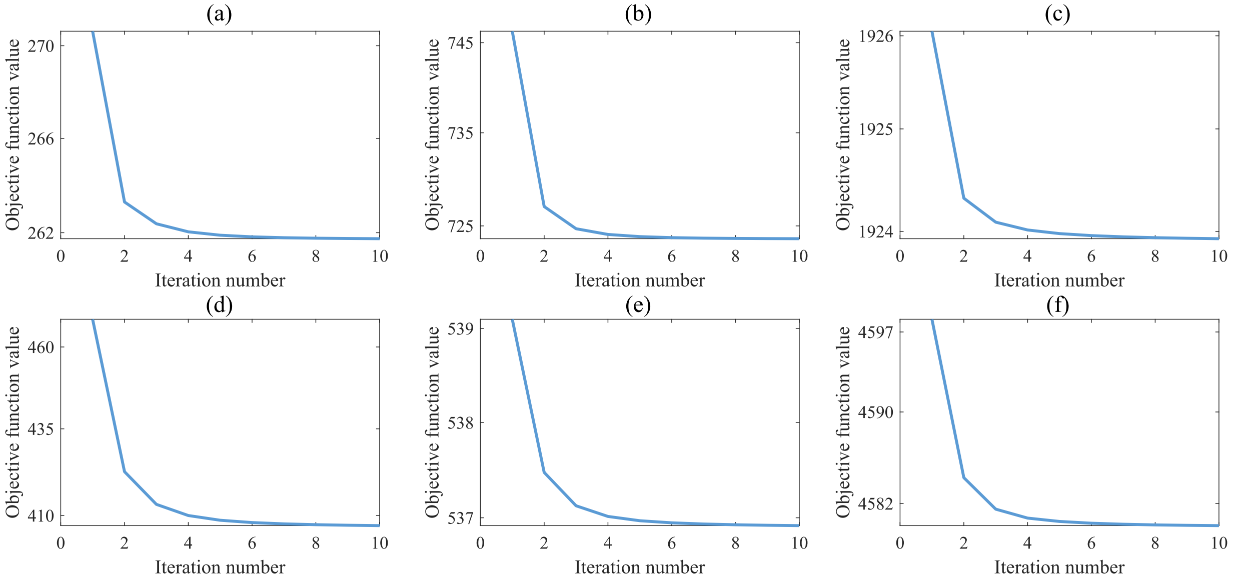


Fig. 12. Convergence curve on different datasets. (a) *DD*; (b) *AWA*; (c) *ILSVRC*; (d) *VOC*; (e) *Cifar*; and (f) *SUN*.

(2) Parameter β controls the adjustment degree of the feature discrimination regularization. The feature correlation influence is dominant when the β value is several orders of magnitude larger than the α value, which limits the multi-granularity knowledge coordination. The β value increase leads to poor FS-MGKC performance on most datasets when $\alpha < 10$. This effect indicates that the main focus should lead on discriminating class-specific features to remove redundancy. The FS-MGKC performance is improved by appropriately reducing the β value when $\alpha > 10$.

The above-mentioned results show that FS-MGKC is sensitive to parameter variation. Therefore, FS-MGKC selects the most discriminative features on different classification tasks by adjusting the parameters.

4.9. Convergence analysis

We discuss and analyze the FS-MGKC convergence on partial hierarchical datasets. Fig. 12 shows the convergence curve of FS-MGKC, where the vertical and horizontal axes mean the convergence function value and iteration number. The iteration number of FS-MGKC is 10 on different datasets, and the objective function of the FS-MGKC convergence curve is the same as Eq. (8).

From Fig. 12, we obtain that the objective function value of FS-MGKC has different convergence values for different datasets and moves in the convergence direction. As the iteration number gradually increases, the objective function values of different datasets monotonically decrease until converge to different stable values.

5. Conclusions and future work

We propose feature selection based on structural manifold learning with multi-granularity knowledge coordination (FS-MGKC). We use the graph Laplacian to combine class dependencies (including strong and weak link relationships), as the manifold coordination regularization term. This automatically coordinates granularity knowledge across multiple levels, which positively influences the perception of the local class structure. The sparsity and correlation between the features are further considered to remove redundancy as the feature differentiation regularization term. Therefore, FS-MGKC selects features highly task-related and with low redundancy.

To reduce the difficulty of large-scale DAG tasks, alternative methods or modifications may be required for performing large-scale classification tasks with DAG structures. We will explore scalable graph clustering algorithms, optimize computational efficiency, or develop new methods specifically designed for large-scale DAG classification.

CRediT authorship contribution statement

Jie Shi: Conceptualization, Methodology, Software, Writing – original draft. **Hong Zhao:** Conceptualization, Methodology, Supervision, Validation, Writing – review & editing.

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.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant Nos. 62376114 and 62141602 and the Natural Science Foundation of Fujian Province under Grant Nos. 2021J011003 and 2021J02049.

References

[1] S. Guo, H. Zhao, Hierarchical classification with multi-path selection based on granular computing, Artif. Intell. Rev. 54 (2021) 2067–2089.

[2] Z. Jiang, K. Liu, X. Yang, H. Yu, H. Fujita, Y. Qian, Accelerator for supervised neighborhood based attribute reduction, Int. J. Approx. Reason. 119 (2020) 122–150.

[3] Y. Chen, P. Wang, X. Yang, J. Mi, D. Liu, Granular ball guided selector for attribute reduction, Knowl.-Based Syst. 229 (2021) 107326.

[4] W. Shu, J. Yu, Z. Yan, W. Qian, Semi-supervised feature selection for partially labeled mixed-type data based on multi-criteria measure approach, Int. J. Approx. Reason. 153 (2023) 258–279.

[5] K. Liu, X. Yang, H. Yu, H. Fujita, X. Chen, D. Liu, Supervised information granulation strategy for attribute reduction, Int. J. Mach. Learn. Cybern. 11 (2020) 2149–2163.

[6] X. Rao, X. Yang, X. Yang, X. Chen, D. Liu, Y. Qian, Quickly calculating reduct: an attribute relationship based approach, Knowl.-Based Syst. 200 (2020) 106014.

[7] X. Liu, Y. Zhou, H. Zhao, Robust hierarchical feature selection driven by data and knowledge, Inf. Sci. 551 (2021) 341–357.

[8] J. Shi, Z. Li, H. Zhao, Feature selection via maximizing inter-class independence and minimizing intra-class redundancy for hierarchical classification, Inf. Sci. (2023) 1–18.

[9] H. Zhao, Q. Hu, P. Zhu, Y. Wang, P. Wang, A recursive regularization based feature selection framework for hierarchical classification, IEEE Trans. Knowl. Data Eng. 33 (7) (2021) 2833–2846.

[10] Y. Lin, Q. Hu, J. Liu, X. Zhu, X. Wu, MULFE: multi-label learning via label-specific feature space ensemble, Trans. Knowl. Discov. Data 16 (2021) 1–12.

[11] Z. Noorie, F. Afsari, Using sparse learning for feature selection with locality structure preserving based on positive data, in: Congress on Fuzzy and Intelligent Systems, 2018, pp. 50–53.

[12] S. Hijazi, D. Hamad, M. Kalakech, A. Kalakech, Active learning of constraints for weighted feature selection, Adv. Data Anal. Classif. 15 (2) (2021) 337–377.

[13] Z. Cai, W. Zhu, Multi-label feature selection via feature manifold learning and sparsity regularization, Int. J. Mach. Learn. Cybern. 9 (8) (2018) 1321–1334.

[14] J. Zhang, Z. Luo, C. Li, C. Zhou, S. Li, Manifold regularized discriminative feature selection for multi-label learning, Pattern Recognit. 95 (2019) 136–150.

[15] J. Hu, Y. Li, W. Gao, P. Zhang, Robust multi-label feature selection with dual-graph regularization, Knowl.-Based Syst. 203 (2020) 106126.

[16] Q. Yin, J. Zhang, S. Wu, H. Li, Multi-view clustering via joint feature selection and partially constrained cluster label learning, Pattern Recognit. 93 (2019) 380–391.

[17] M. Rostami, K. Berahmand, S. Forouzandeh, A novel method of constrained feature selection by the measurement of pairwise constraints uncertainty, J. Big Data 7 (1) (2020) 1–21.

[18] Y. Wang, L. Chen, J. Zhou, T. Li, Y. Yu, Pairwise constraints-based semi-supervised fuzzy clustering with multi-manifold regularization, Inf. Sci. 638 (2023) 118994.

[19] I. Jo, S. Lee, S. Oh, Improved measures of redundancy and relevance for mRMR feature selection, Computers 8 (2) (2019) 42–55.

[20] H. Lim, D. Kim, Pairwise dependence-based unsupervised feature selection, Pattern Recognit. 111 (2021) 107663.

[21] R. Tibshirani, Regression shrinkage and selection via the lasso, J. R. Stat. Soc. 58 (1) (1996) 267–288.

[22] Z. Wang, D. Wu, R. Wang, F. Nie, F. Wang, Joint anchor graph embedding and discrete feature scoring for unsupervised feature selection, IEEE Trans. Neural Netw. Learn. Syst. (2022) 1–14.

[23] P. Huang, Z. Kong, M. Xie, X. Yang, Robust unsupervised feature selection via data relationship learning, Pattern Recognit. 142 (2023) 109676.

[24] C. Li, H. Li, Network-constrained regularization and variable selection for analysis of genomic data, Bioinformatics 24 (9) (2008) 1175–1182.

[25] G. Miller, WordNet: a lexical database for English, Commun. ACM 38 (11) (1995) 39–41.

[26] Y. Zhang, J. Chen, Z. Liu, Adaptive distance penalty based nonnegative low-rank representation for semi-supervised learning, Appl. Intell. 53 (2) (2023) 1405–1416.

[27] J. Chavoshinejad, S. Seyedi, F. Tab, N. Salahian, Self-supervised semi-supervised nonnegative matrix factorization for data clustering, Pattern Recognit. 137 (2023) 109282.

[28] M. Yu, Y. Zhou, R. Li, X. Wang, Y. Zhong, Semi-supervised learning via manifold regularization, J. China Univ. Post Telecommun. 19 (6) (2012) 79–88.

[29] Y. Lin, H. Liu, H. Zhao, Q. Hu, X. Zhu, X. Wu, Hierarchical feature selection based on label distribution learning, IEEE Trans. Knowl. Data Eng. (2022) 5964–5976.

[30] J. Gan, G. Wen, H. Yu, W. Zheng, C. Lei, Supervised feature selection by self-paced learning regression, Pattern Recognit. Lett. 132 (2020) 30–37.

[31] A. Argyriou, T. Evgeniou, M. Pontil, Multi-task feature learning, Adv. Neural Inf. Process. Syst. 19 (2006) 41–48.

[32] I. Dimitrovski, D. Kocev, S. Loskovska, S. Deroski, Hierarchical annotation of medical images, Pattern Recognit. 44 (10–11) (2011) 2436–2449.

[33] C. Lampert, H. Nickisch, S. Harmeling, Learning to detect unseen object classes by between-class attribute transfer, in: Conference on Computer Vision and Pattern Recognition, 2009, pp. 951–958.

[34] M. Everingham, L. VanGool, C. Williams, J. Winn, A. Zisserman, The Pascal visual object classes (VOC) challenge, Int. J. Comput. Vis. 88 (2) (2010) 303–338.

[35] A. Krizhevsky, G. Hinton, Learning multiple layers of features from tiny images, Handb. Syst. Autoimmune Dis. 1 (4) (2009) 1–60.

[36] J. Krause, M. Stark, J. Deng, L. Feifei, 3D object representations for fine-grained categorization, in: IEEE Conference on Computer Vision Workshops, 2013, pp. 554–561.

[37] J. Xiao, J. Hays, K. Ehinger, A. Oliva, A. Torralba, SUN database: large-scale scene recognition from abbey to zoo, in: Conference on International Journal of Computer Vision, 2016, pp. 3–22.

[38] C. Ding, I. Dubchak, Multi-class protein fold recognition using support vector machines and neural networks, Bioinformatics 17 (4) (2001) 349–358.

[39] L. Wei, M. Liao, X. Gao, Q. Zou, An improved protein structural classes prediction method by incorporating both sequence and structure information, IEEE Trans. Nanobiosci. 14 (4) (2014) 339–349.

[40] M. Ashburner, C. Ball, J. Blake, D. Botstein, H. Butler, J. Cherry, A. Davis, K. Dolinski, S. Dwight, J. Eppig, Gene ontology: tool for the unification of biology, Nat. Genet. 25 (1) (2000) 25–29.

[41] P. Hart, D. Stork, R. Duda, Pattern Classification, Wiley, Hoboken, 2000.

[42] F. Nie, H. Huang, X. Cai, C. Ding, Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization, Adv. Neural Inf. Process. Syst. 23 (2010) 1813–1821.

[43] L. Grimaudo, M. Mellia, E. Baralis, Hierarchical learning for fine grained Internet traffic classification, in: Conference on Wireless Communications and Mobile Computing, 2012, pp. 463–468.

[44] X. Liu, H. Zhao, Robust hierarchical feature selection with a capped ℓ_2-norm, Neurocomputing 443 (2021) 131–146.

[45] Q. Tuo, H. Zhao, Q. Hu, Hierarchical feature selection with subtree based graph regularization, Knowl.-Based Syst. 163 (2019) 996–1008.

[46] Y. Lin, S. Bai, H. Zhao, S. Li, Q. Hu, Label-correlation-based common and specific feature selection for hierarchical classification, J. Softw. 33 (7) (2022) 2667–2682.

[47] L. Cai, T. Hofmann, Exploiting known taxonomies in learning overlapping concepts, in: Conference on Artificial Intelligence, 2007, pp. 708–713.

[48] B. Schieber, U. Vishkin, On finding lowest common ancestors: simplification and parallelization, SIAM J. Comput. 17 (6) (1988) 1253–1262.

[49] O. Dekel, J. Keshet, Y. Singer, Large margin hierarchical classification, in: Conference on Machine Learning, vol. 27, 2004.

[50] O. Dunn, Multiple comparisons among means, J. Am. Stat. Assoc. 56 (293) (1961) 52–64.