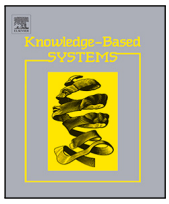
Knowledge-Based Systems 314 (2025) 113210

Contents lists available at ScienceDirect

Knowledge-Based Systems

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



Fine-grained local label correlation for multi-label classification

Tianna Zhao [a,b,c], Yuanjian Zhang [d,*], Duoqian Miao [e,f], Witold Pedrycz [g,h,i]

[a] Institute of Artificial Intelligence on Education, Shanghai Normal University, 200234, China
[b] The Research Base of Online Education for Shanghai Middle and Primary Schools, Shanghai Normal University, 200234, China
[c] Shanghai Engineering Research Center of Intelligent Education and Big data, Shanghai Normal University, 200234, China
[d] School of Computer Engineering and Science, Shanghai University, 200444, China
[e] Department of Computer Science and Technology, Tongji University, Shanghai, 201804, China
[f] Key Laboratory of Embedded System and Service Computing, Ministry of Education, Tongji University, Shanghai, 201804, China
[g] Silesian University of Technology (SUT), Department of Measurement and Control Systems,, Gliwice, Akademicka 2, 44-100, Poland
[h] Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB T6G 2R3, Canada
[i] Research Center of Performance and Productivity Analysis, Istinye University, Istanbul, Türkiye

ARTICLE INFO

Keywords:
Multi-granularity
Feature augmentation
Local label correlation
Label-specific features
Multi-label classification

ABSTRACT

Comprehensive learning label correlation is conducive to boosting the accuracy of multi-label classification. While existing methods focus on exploring the correlation-aware original features or latent subspaces, they often overlook the role of correlation in deducing local structures. The oversight can result in suboptimal topic-based label correlation estimation and thus incur information loss. In contrast to the conventional single-granularity-based learning for local label correlation, we propose a multi-granularity correlation-based feature augmentation (MGOFA) model. MGOFA consists of three components that progressively refine the granularity of label correlation: granular-based feature augmentation for relative neighborhood-based class tendency estimation, granular-based latent topic mining for tendency-aware topic modeling, and fine-grained label correlation mining for augmented local label correlation learning. The information on neighborhood-based similarity between instances is explicitly leveraged and contributes to the model two-fold. Firstly, it induces the prototypes of latent topics, which share more knowledge with the label association. Secondly, it refines the discriminative granularity of the model by integrating it with the original features. Such a formulation simulates the viewpoint of human decision-making by automatically determining optimal solutions on both data and knowledge from coarse and refined granularity, respectively. Extensive comparisons completed of ten benchmarks demonstrate that MGOFA outperforms the state-of-the-art methods with satisfying convergence and sensitivity.

1. Introduction

Multi-label classification [1,2] determines the label associations of instances by learning a projection from features to labels. The instances may be associated with varying collections of labels, which can be co-existent or mutually exclusive. The label association described by a probability distribution is an instantiation of label correlation. For example, the labels *beach*, *boat*, *harbor* can be available in an image, while the *sun* and *star* are less likely to appear together. Such phenomena associated with labels imply the topic distribution and are omnipresent in real applications involving emotion analysis [3], video annotation [4], and disease diagnosis [5].

Label correlation is an indispensable component in improving the model generalization. From unconditionally hold only to both conditionally and unconditionally hold, there are three categories called *global label correlation*, *local label correlation*, and *global and local label correlation*. The *global label correlation* corresponds to the case where label correlation holds unconditionally at the same possibility across the instances. This assumption favors those topic-independent label correlations. However, it incurs performance degeneration due to the over-simplification of topic structures. Representatives include Random *k*-label sets [6], LIFT [7], LLSF [8], and HNOML [9]. In contrast, the *local label correlation* embraces the case where both strengths and components are topic-dependent. Representatives include ML-LOC [10], GD-LDL-SCL [11], LTE [12], and MLCD [13]. The third group (i.e., *global and local label correlations*) is an integration of the first two assumptions and attempts to reach a balance between topic-dependent and topic-independent structures. Representatives include Glocal [14], MDFS [15], TIFS [16], and MLC-LFLC [17]. Considering

* Corresponding author.
E-mail address: zhangyuanjian@shu.edu.cn (Y. Zhang).

https://doi.org/10.1016/j.knosys.2025.113210
Received 23 June 2024; Received in revised form 1 January 2025; Accepted 18 February 2025
Available online 27 February 2025

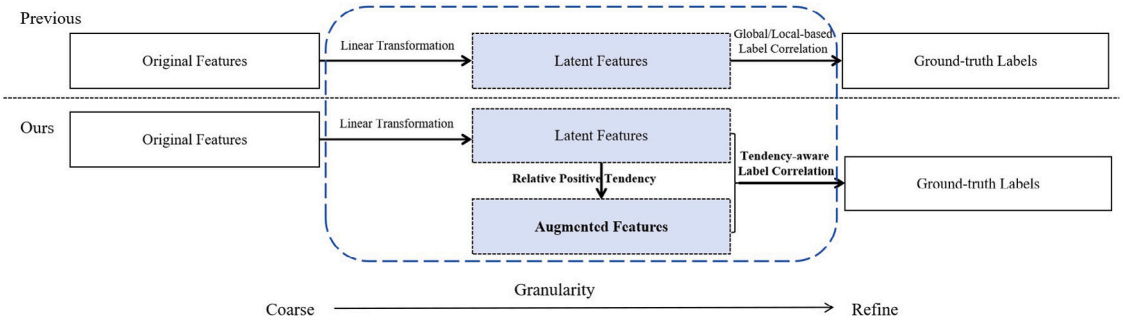


Fig. 1. Major differences between existing approaches and our approach.

that the diversified topics are a generalization case of a holistic topic, it is necessary to introduce a novel decision-making theory to reduce the uncertainty of topics by approximating the underlying structures of label correlation.

Granular computing(GrC) [18,19] is a structural cognitive methodology that simulates human processing via measuring and reasoning on knowledge characterized by information granules. Many scholars [20–25] demonstrate the superiority of employing granular computing on multi-label classification. With granular computing, it is likely to optimize the label association by approximating the underlying structures of label correlation via the refinements on granules. The construction of information granules can explicitly describe the correlations between instances and labels by combining them with heterogeneous techniques. Fortunately, the integrations of the neighborhood, clustering, and feature augmentation technique [26,27] are competent to complete such a task. To realize the adaptive utilization of multi-faceted information, we devise a correlation-based multi-granulation feature augmentation model to hierarchically approximate the local intrinsic relationship within instances, which assumes a different perspective against the previous solutions (see Fig. 1).

Our contributions are enumerated below:

1. We have developed a multi-granularity structure to highlight the functionality of local label correlation for multi-label classification. The structure sequentially refines the knowledge granularity on label correlation by exploring similarity in correlation-based feature augmentation.
2. The aim of correlation-based feature augmentation is not to directly replace the original feature representation. Instead, it works in conjunction with the original features to enrich the feature representation, allowing for the simulation of the adaptive perception mechanism of humans via optimization.
3. We extend the scope of multi-granularity on multi-label classification by introducing a novel model called multi-granularity correlation-based feature augmentation (MGOFA). Furthermore, we demonstrate that the proposed second-order-based local label correlation not only dominates the state-of-the-art approaches with high-order or sophisticated label correlation but also produces promising performance with respect to convergence and sensitivity.

The paper is organized as follows: Section 2 reviews the related work. Section 3 outlines the pipeline and explains model details. Section 4 presents experimental results on benchmarks and further analyses on the MGOFA. Finally, Section 5 concludes the paper and identifies the future directions.

2. Related work

We briefly review some related work regarding three aspects, i.e., label correlation, multi-granularity, and feature augmentation.

2.1. Label correlation

Label correlation constitutes a widespread assumption in dealing with multi-label classification as they can condense the solution from a sparse distribution of labels. However, one visible challenge is the inadequate knowledge of the label correlation, which can be regarded as the cartesian product of components (which describes how labels are correlated) and scopes (which describes the collections of instances that share the same components of label correlations) on label space (see Fig. 2). The components are relevant to a particular order of label correlation and are generally composed of three categories named as *first-order*, *second-order*, and *high-order*, respectively. In contrast, the scopes are pertinent to the cardinality of topics and are generally composed of three categories named as *global label correlation*, *local label correlation*, and *global and local label correlation*. Fig. 2 enumerates three representative taxonomies with the corresponding representative algorithms to determine whether association of instance-label pairs hold (denoted as 1) or not (denoted as 0). The *second-order* with *global label correlation* include LLSF [8], HNOML [9], and WRAP [28], where Huang et al. [8] exploited label-specific features by imposing ℓ_1-norm on the linear weights of observable features, Zhang et al. [9] enhanced the generalization of second-order label correlation by constructing a projection from feature vectors to enriched labels, Yu et al. [28] established label-specific features in embedded space by introducing ℓ_2-norm on projection and ℓ_1-norm on linear model. The *second-order* with *local label correlation* include LPLC [29], LF-LPLC [30], and GD-LDL-SCL [11], where Huang et al. [29] explored neighborhood-based pairwise label correlation from both the positive and negative perspectives, Weng et al. [30] leveraged clustering-based pairwise label correlation by improving feature discrimination on aligned labels, Jia et al. [11] learned clustering-based pairwise label correlation by imposing ℓ_2-norm on similarity regularization between instances and cluster centers. The *second-order* with *global and local label correlation* include Glocal [14], MDFS [15], and TIFS [16], where Zhu et al. [14] pioneered the global and local label correlation by introducing instance-level and label-level manifold regularization constraints simultaneously, Zhang et al. [15] exploited the global and local label correlation by considering the $\ell_{2,1}$-norm regularization on embedded feature space, Ma et al. [16] captured the topic-based correlation on latent representations constructed by nonnegative matrix factorization.

2.2. Multi-granularity analysis

Multi-granularity [31] of information is a structural methodology of approximate modeling with the uncertainty that highlights the interactions between data and knowledge via the generation and deduction of information granules across multiple layers. The uncertainty measure is a fundamental component in describing the degrees of functionality similarity. For instance, Xia et al. [32] preserved the variation monotonicity of rough approximations in attribute reduction by defining local knowledge distance. Wang et al. [33] formulated the multi-granularity decision problem for large-scale trustworthy hierarchical

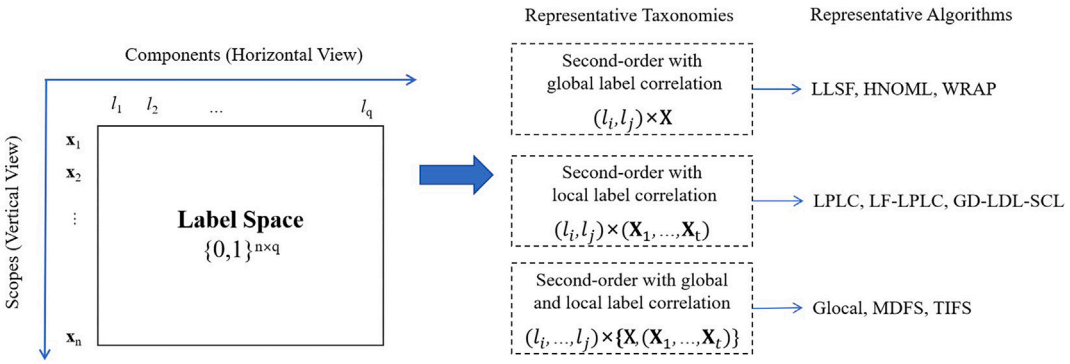


Fig. 2. Representative learning frameworks on label correlation.

Table 1
Differences between MGOFA and the related algorithms.

Methods	Global label correlation	Local label correlation	Feature augmentation	Wrapping feature selection	Embedding feature selection
MGOFA	×	✓	✓	✓	×
WRAP [28]	✓	×	×	✓	×
HOMI [43]	✓	×	×	×	✓
SLOFS [44]	✓	×	×	×	✓
MDFS [15]	✓	✓	×	×	✓
TIFS [16]	✓	✓	×	×	✓
RLFSCL [45]	✓	×	×	×	✓
GLFS [46]	✓	✓	×	×	✓
MC-GM [47]	✓	✓	×	×	✓

✓ denotes that a method takes the corresponding strategy.
× denotes that a method does not take the corresponding strategy.

classification by developing a distortion measure on reconstruction error. Zhang et al. [34] improved the approximation robustness of the double-quantitative neighborhood rough set by optimizing neighborhood distance with arithmetic-mean and geometric-mean, respectively. Shu et al. [35] renewed the hierarchical structures of fine-grained image-based classification by defining joint probability-based loss. Some recent advancements in multi-label classification manifest the effectiveness. For instance, Yu et al. [36] enriched the inference capability of graph convolutional neural networks by introducing attribute subset-based object relationships. Yu et al. [37] emphasized the uncertainty of inter-feature relationships in boundary regions by characterizing instance-level tripartition based on the variable degree of the multi-granulation rough set.

2.3. Feature augmentation

Feature augmentation [38] is an essential solution to enrich the discrimination of original features by generating informative features with the semantics of class-oriented correlation. It is different from the well-established feature selection techniques (e.g., filtering, wrapping, and embedding) strengthen the feature-label correlation by replacing rather than replenishing the original. Recently, feature augmentation has demonstrated effectiveness in the single-label classification cases [39–42] and multi-dimensional classification cases [26,27]. Zhang et al. [39] constructed a feature enhancement network with characteristics of multi-granularity to improve the classification performance of image-based small object detection. Wang et al. [41] incorporated feature augmentation in the contrastive semantic augmentation loss to alleviate the negative transfer. Jia et al. [26] first introduced correlation-based feature augmentation to the multi-dimensional classification case to address the inappropriate fitness on class dependencies. As multi-dimensional classification is an extension of multi-label classification, we believe it can be a powerful module to boost classification performance. The differences between MGOFA and the related state-of-the-art multi-label classification methods are summarized in Table 1.

3. Proposed model

3.1. Notation

Given a multi-label dataset $D = \{(\mathbf{x}_i, Y_i) \mid 1 \leqslant i \leqslant n\}$, let $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]^\top \in \mathbb{R}^{n \times m}$ and $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n]^\top \in \{0,1\}^{n \times q}$ denote the n instances with m-dimensional features and q-dimensional labels on n instances, where $\mathbf{y}_i = [y_{i1}, y_{i2}, \ldots, y_{iq}]^\top$. $y_{ij} = 1$ if $l_j \in Y_i$ and $y_{ij} = 0$ otherwise. MGOFA generates the topic-based label-specific features by learning correlation similarity from observable features. For ease of reference, we enumerate major notation in Table 2.

3.2. Basic idea

We optimize the label-specific feature representations by learning discriminative yet limited features in a global-to-local strategy. As described in Fig. 3, MGOFA includes three crucial components called "Granular-based Feature Augmentation", "Granular-based Latent Topic Mining", and "Fine-grained Label Correlation Learning", respectively. To approximate the latent topic-aware label correlation, we first enrich the instance correlation by learning augmented features with q-dimensionality in a label-by-label manner. For the jth label l_j, the augmented feature $\mathbf{AF}_j \in \mathbb{R}^{n \times 1}$ measures the relative positive class tendency on l_j. The feature prototype of l_j learns from the instance-based neighborhood determined by trade-off factor μ and neighborhood count k simultaneously. By exploring the instance and class distribution within the neighborhood, we refine the prototypes with the overall weighted scores for positive and negative (see $\mathbf{WSP}_j$ and $\mathbf{WSN}_j$ in Fig. 3) and deduce the augmented features accordingly. Secondly, we explore the underlying topic distribution by concatenating augmented features across all labels rather than the original features. Thirdly, we partition the original features and the corresponding augmented features into different topics (exemplified as $\mathbf{X}^1_{aug}$, $\mathbf{X}^2_{aug}$, and $\mathbf{X}^3_{aug}$ in Fig. 3) by generating multi-granularity features. With the original features (abbreviated as OF) and augmented features

Table 2
Summary of notations.

Notations	Meanings
$\mathbf{X} \in \mathbb{R}^{n\times m}$	Original training set with n instances and m features
$\mathbf{Y} \in \mathbb{R}^{n\times q}$	Labelset of $\mathbf{X}$ on q labels
$Dis(\cdot,\cdot)$	Instance-based distance operator
μ	Weight of cosine similarity and Pearson correlation coefficient in instance-based distance
k	Size of nearest neighbor $\forall \mathbf{x}_b$
$\mathcal{I}(\mathbf{x}_b)$	An ordered indicator of the k nearest neighbors of $\mathbf{x}_b$ based on the ascending order of $Dis(\cdot,\cdot)$
$\mathcal{P}_j^{\mathbf{x}_b} \in \mathbb{R}^{k\times 1}$	Indicators of the nearest neighborhood with label association on the jth label for $\mathbf{x}_b$
$\mathcal{N}_j^{\mathbf{x}_b} \in \mathbb{R}^{k\times 1}$	Indicators of the nearest neighborhood without label association on the jth label for $\mathbf{x}_b$
$\mathbf{W}^{\mathbf{x}_b} \in \mathbb{R}^{k\times 1}$	Instance weights of k nearest neighborhood for $\mathbf{x}_b$
$\mathbf{rip}_j \in \mathbb{R}^{n\times 1}$	Relative influential ratio of positive instances within $\mathcal{I}(\mathbf{x}_b)$
$\mathbf{rin}_j \in \mathbb{R}^{n\times 1}$	Relative influential ratio of negative instances within $\mathcal{I}(\mathbf{x}_b)$
$\mathbf{WSP}_j \in \mathbb{R}^{n\times 1}$	Weighted score vectors for positive tendency of instances on label l_j
$\mathbf{WSN}_j \in \mathbb{R}^{n\times 1}$	Weighted score vectors for negative tendency of instances on label l_j
$Me(\mathbf{W}^{\mathbf{x}_b})$	Median radius of the k nearest neighborhood w.r.t. $\mathbf{x}_b$
$\mathbf{AF}_j \in \mathbb{R}^{n\times 1}$	Augmented feature for label l_j
$\mathbf{X}_{aug} \in \mathbb{R}^{n\times(m+q)}$	Multigranularity features
t	Latent topic count
$\mathbf{X}_{aug}^r \in \mathbb{R}^{n,\times(m+q)}$	The feature representation of the rth topic
$\mathbf{V}_r \in \mathbb{R}^{(m+q)\times d}$	Embedding matrix for generation of multi-granularity features w.r.t. the rth topic
$\mathbf{U}_r \in \mathbb{R}^{d\times q}$	Weight matrix for multi-granularity features w.r.t. the rth topic
$\mathbf{b}_r \in \mathbb{R}^{q\times 1}$	Bias vector for multi-granularity features w.r.t. the rth topic
$E(\cdot)$	Expectation of a variable
$Bias_o(\mathbf{X})$	Bias term incurred by the original features without consideration of topics
$Bias_c(\mathbf{X}_{aug})$	Bias term incurred by concatenation of original and augmented features without consideration of topics
$Bias_m^r(\mathbf{X}_{aug}^r)$	Bias term incurred by latent topics from multi-granularity features

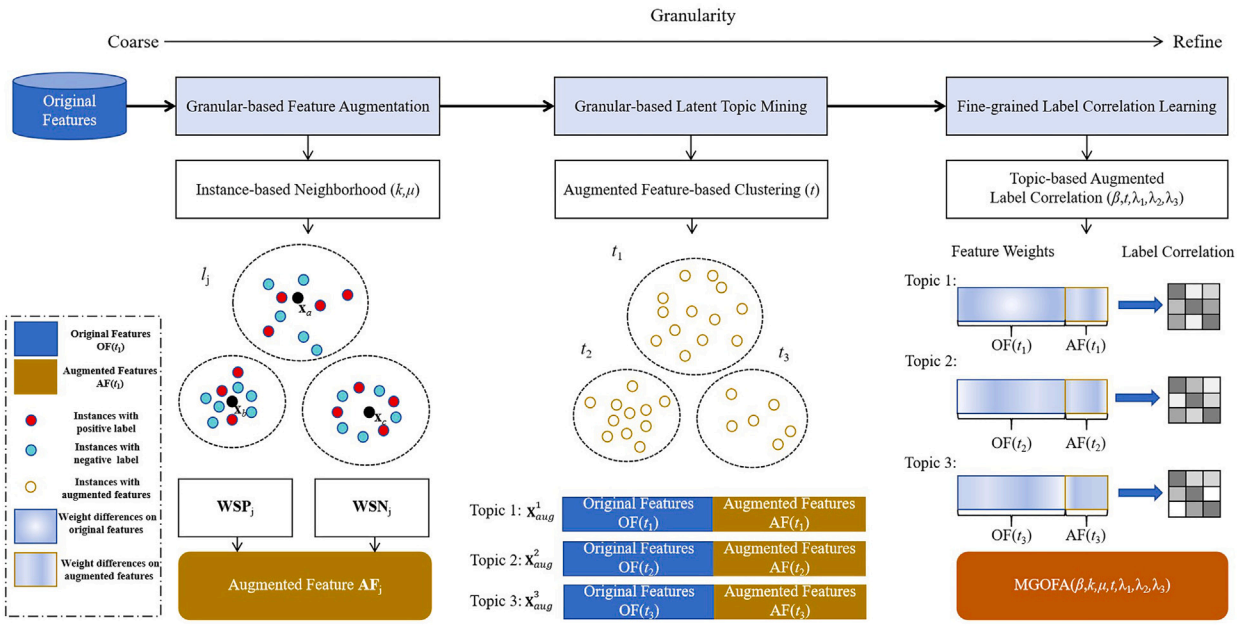


Fig. 3. Framework of MGOFA.

(abbreviated as AF), we optimize the topic-based label correlation by learning the local concatenation of OF and AF in a wrapped strategy. Ultimately, we obtain a fine-grained solution of weight combinations on original and augmented features.

3.3. Granular-based feature augmentation

3.3.1. Instance-based neighborhood

The instance-based neighborhood measures the likelihood of the consistent label association between two arbitrary instances. Here, we construct a granular-based neighborhood determined by parameters k and μ simultaneously. For an arbitrary instance $\mathbf{x}_b$, let $\mathcal{I}(\mathbf{x}_b) = \left(Dis\left(\mathbf{x}_b, \mathbf{x}_{(1)}\right), \ldots, Dis\left(\mathbf{x}_b, \mathbf{x}_{(k)}\right)\right)$ denote the ordered set of indexes (except the instance $\mathbf{x}_b$ itself) identified in k-nearest neighborhood of $\mathbf{x}_b \in \mathbf{X}$ based on $Dis(\cdot,\cdot)$, which is a mixture of Pearson correlation coefficient and cosine similarity and determined by $\mathbf{x}_b(k,\mu)$.

$$Dis(\mathbf{x}_b,\mathbf{x}_r) = \frac{\mu}{2}\left(1 - \frac{Cov(\mathbf{x}_b,\mathbf{x}_r)}{\sigma_{\mathbf{x}_b}\sigma_{\mathbf{x}_r}}\right) + \frac{1-\mu}{2}\left(1 - \frac{\mathbf{x}_b\mathbf{x}_r}{\|\mathbf{x}_b\|\|\mathbf{x}_r\|}\right) \tag{1}$$

where $\frac{Cov(\mathbf{x}_b,\mathbf{x}_r)}{\sigma_{\mathbf{x}_b}\sigma_{\mathbf{x}_r}}$ and $\frac{\mathbf{x}_b\mathbf{x}_r}{\|\mathbf{x}_b\|\|\mathbf{x}_r\|}$ represent Pearson correlation coefficient and cosine similarity, respectively. Note that Pearson correlation coefficient and cosine similarity share similar values if the linear correlation

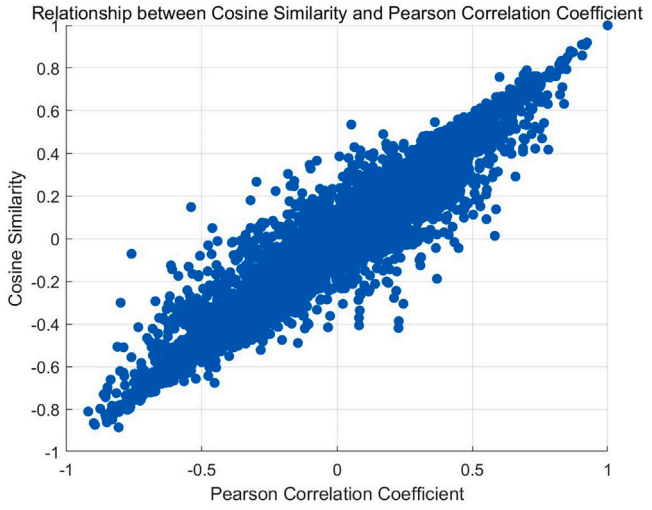


Fig. 4. Visualization of differences between Pearson correlation coefficient and cosine similarity.

becomes stronger but diversified otherwise (see Fig. 4). It is thus meaningful to integrate them by introducing a trade-off factor μ. Typically, $Dis(\cdot,\cdot)$ assumes values in the unit interval as presented in Property 1.

Property 1. $Dis(\cdot,\cdot) \in [0,1]$

Proof. Since $\frac{Cov(\mathbf{x}_b,\mathbf{x}_r)}{\sigma_{\mathbf{x}_b}\sigma_{\mathbf{x}_r}} \in [-1,1]$ and $\frac{\mathbf{x}_b\mathbf{x}_r}{\|\mathbf{x}_b\|\|\mathbf{x}_r\|} \in [-1,1]$ hold, which means both $\frac{\mu}{2}\left(1-\frac{Cov(\mathbf{x}_b,\mathbf{x}_r)}{\sigma_{\mathbf{x}_b}\sigma_{\mathbf{x}_r}}\right) \in [0,2]$ and $\frac{1-\mu}{2}\left(1-\frac{\mathbf{x}_b\mathbf{x}_r}{\|\mathbf{x}_b\|\|\mathbf{x}_r\|}\right) \in [0,2]$ hold. Considering that $\mu \in [0,1]$, we have $Dis(\cdot,\cdot) \in [0,1]$. □

As the value of $Dis(\cdot,\cdot)$ decreases, the similarity of the $\mathbf{x}_r$ for the specified $\mathbf{x}_b$ increases, indicating a stronger similarity in the label association between $\mathbf{x}_b$ and $\mathbf{x}_r$.

3.3.2. Multi-layer feature augmentation

We conduct feature augmentation by estimating neighborhood-based class tendencies for each label. As positive and negative are complementary, we estimate positive tendency by considering the following factors:

1. (Coarse granules): The percentage of instances with the positive class within the neighborhood of $\mathbf{x}_b$.
2. (Refined granules): The relative closeness of instances with the positive class within the neighborhood of $\mathbf{x}_b$.
3. (Hybrid granules): The compactness of k neighborhood of $\mathbf{x}_b$.

The positive tendency of $\mathbf{x}_b$ on l_j increases as the values of coarse, refined, and hybrid granules become larger. Here, we offer some explanations of rationality. Specifically, a larger ratio of instances with a positive class on l_j in the neighborhood of $\mathbf{x}_b$ means that $\mathbf{x}_b$ is surrounded by more instances that are associated with l_j, thus $\mathbf{x}_b$ is more likely to be associated with l_j. However, a drawback is that the coarse granules treat all relevant instances equally. A more reasonable solution is to examine the values on refined granules, which assumes that the instances contribute more to class tendency determination if they are similar. It is worth mentioning that the combination of both coarse and refined granules is not comprehensive, as the radius of k nearest neighborhood can be very different. Consequently, the hybrid granule measures the reliability of the combination objectively.

We consider both instance similarity weight and relative influential ratio on refined granules. The instance similarity weight measures how many degrees of similarity for an arbitrary instance $\mathbf{x}_r$ to the given instance $\mathbf{x}_b$, which can be deduced from distance measure $Dis(\mathbf{x}_b,\mathbf{x}_r)$. Formally, we devise a similar-aware weight vector (i.e., $\mathbf{W}^{\mathbf{x}_b}$) in computing varying $\mathcal{I}(\mathbf{x}_b)$.

$$\mathbf{W}^{\mathbf{x}_b} = \left(\sqrt{1-Dis(\mathbf{x}_b,\mathbf{x}_r)}\right)_{k\times 1} \tag{2}$$

where $\mathbf{x}_r$ is a component of k nearest neighborhood of $\mathbf{x}_b$. It can be easily deduced the component $w_r^{\mathbf{x}_b} \in [0,1]$, with the larger weight of $\mathbf{x}_r$ if more similar to $\mathbf{x}_b$ and vice versa. Example 1 explains the rationality of Eq. (2).

Example 1. Let $\mathbf{x}_b$ and $\mathbf{x}_c$ be two instances, given that $(k,\mu) = (5, 0.5)$ and two ordered indicators denoted as $Dis(\mathbf{x}_b, \mathcal{I}(\mathbf{x}_b)) = (0.10, 0.20, 0.25, 0.30, 0.50)$ and $Dis(\mathbf{x}_b, \mathcal{I}(\mathbf{x}_c)) = (0.20, 0.30, 0.45, 0.50, 0.70)$. Then the $\mathbf{W}^{\mathbf{x}_b}$ and $\mathbf{W}^{\mathbf{x}_c}$ are as follows:

$$\mathbf{W}^{\mathbf{x}_b} = \left(\sqrt{0.90}, \sqrt{0.80}, \sqrt{0.75}, \sqrt{0.70}, \sqrt{0.50}\right)_{5\times 1}$$

$$\mathbf{W}^{\mathbf{x}_c} = \left(\sqrt{0.80}, \sqrt{0.70}, \sqrt{0.55}, \sqrt{0.50}, \sqrt{0.30}\right)_{5\times 1}$$ □

Obviously, Eq. (2) offers a data-driven manner to adaptively adjust the influential strength of the neighborhood to the center instance. Typically, $w_r^{\mathbf{x}_b}$ reaches maximum value of 1 if the original features of $\mathbf{x}_b$ and $\mathbf{x}_r$ are indistinguishable.

The coarse granule construction are constructed based on the neighborhood of instances. We realize the coarse granule by calculating the instance count with positive/negative class on l_j. Suppose the distance between $\mathbf{x}_b$ and an arbitrary $\mathbf{x}_r$ ranks in ascending order according to $Dis(\cdot,\cdot)$, where $r = 1, 2, \ldots, n$. The indicators $\mathcal{P}_j^{\mathbf{x}_b}$ and $\mathcal{N}_j^{\mathbf{x}_b}$ are defined as:

$$\mathcal{P}_j^{\mathbf{x}_b} = \left(\llbracket y_{r,j} = 1 \rrbracket\right)_{k\times 1} \tag{3}$$

where $\llbracket \pi \rrbracket$ returns 1 if it holds and 0 otherwise. Similarly, we have:

$$\mathcal{N}_j^{\mathbf{x}_b} = \left(\llbracket y_{r,j} = 0 \rrbracket\right)_{k\times 1} \tag{4}$$

where $\llbracket \pi \rrbracket$ returns 1 if it holds and 0 otherwise. The relationship between $\mathcal{P}_j^{\mathbf{x}_b}$ and $\mathcal{N}_j^{\mathbf{x}_b}$ is given in Property 2.

Property 2. $\sum \mathcal{P}_j^{\mathbf{x}_b} + \sum \mathcal{N}_j^{\mathbf{x}_b} = k$

Proof. Eqs. (3) and (4) imply that both $\mathcal{P}_j^{\mathbf{x}_b}$ and $\mathcal{N}_j^{\mathbf{x}_b}$ are composed of the sequence of $\{0,1\}^k$. Since the instance $\mathbf{x}_r$ is either associated with label l_j or not, which means either $\llbracket y_{r,j} = 1 \rrbracket$ or $\llbracket y_{r,j} = 0 \rrbracket$ holds. Therefore, the element-wise summation on $\mathcal{P}_j^{\mathbf{x}_b}$ and $\mathcal{N}_j^{\mathbf{x}_b}$ should be a vector of all 1 with the dimensionality of $1 \times k$. □

Meanwhile, the relative influential ratio measures the similarity between the observed class-based distribution and the optimal distributions of positive and negative classes. Concretely, the optimal distribution of the positive class corresponds to the case where all positive instances within the k neighborhood are the $\sum \mathcal{P}_j^{\mathbf{x}_b}$ nearest. The optimal distribution of the negative is analogous. Formally, we define the relative influential ratio for a positive class l_j as:

$$\mathbf{rip}_j = \left(rip_j^{\mathbf{x}_b}\right)_{n\times 1}$$

$$where \quad rip_j^{\mathbf{x}_b} = \begin{cases} \frac{\mathbf{W}^{\mathbf{x}_b\top}\mathcal{P}_j^{\mathbf{x}_b}}{\sum_{r=1}^{P} w_r^{\mathbf{x}_b}} & \exists \llbracket y_{rj} = 1 \rrbracket \\ 0 & otherwise \end{cases} \tag{5}$$

where $\mathbf{x}_b \in \mathbf{X}$ and $P = \mathbf{1}_k\mathcal{P}_j^{\mathbf{x}_b}$, $\mathbf{1}_k$ denotes a row vector with all elements of length k being 1.

Analogously, we define the relative influential ratio for the negative class on l_j as:

$$\mathbf{rin}_j = \left(rin_j^{\mathbf{x}_b}\right)_{n\times 1}$$

$$where \quad rin_j^{\mathbf{x}_b} = \begin{cases} \frac{\mathbf{W}^{\mathbf{x}_b\top}\mathcal{N}_j^{\mathbf{x}_b}}{\sum_{r=1}^{N} w_r^{\mathbf{x}_b}} & \exists \llbracket y_{rj} = 0 \rrbracket \\ 0 & otherwise \end{cases} \tag{6}$$

where $\mathbf{x}_b \in \mathbf{X}$ and $N = \mathbf{1}_k \mathcal{N}_j^{\mathbf{x}_b}$, $\mathbf{1}_k$ denotes a row vector with all elements of length k being 1.

Example 2 shows the computing of relative influential ratio on $\mathbf{x}_b$ and $\mathbf{x}_c$.

Example 2 (*Continuing Example 1*)**.** Let $y_{\mathbf{x}_b j} = 1$ and $y_{\mathbf{x}_c j} = 1$ denote the instance $\mathbf{x}_b$ and $\mathbf{x}_c$ are positively and negatively associated with l_j. Suppose the instances in $\mathcal{I}(\mathbf{x}_b)$ are all negatively associated with l_j, while the instances with the second-nearest and third-nearest of $\mathbf{x}_c$ are negatively associated with l_j, then the relative influential ratio $rip_j^{\mathbf{x}_b}$ and $rin_j^{\mathbf{x}_c}$ are computed as:

$$rip_j^{\mathbf{x}_b} = 0$$

$$rin_j^{\mathbf{x}_c} = \frac{\sqrt{0.70} + \sqrt{0.55}}{\sqrt{0.80} + \sqrt{0.70}} = 0.9117 \quad \square$$

It should be mentioned that the elements of $\mathbf{rip}_j$ and $\mathbf{rin}_j$ (i.e., $rip_j^{\mathbf{x}_b}$ and $rin_j^{\mathbf{x}_b}$) are finite as demonstrated in Property 3.

Property 3. *For $rip_j^{\mathbf{x}_b}$ and $rin_j^{\mathbf{x}_b}$, we have:*

(1) $rip_j^{\mathbf{x}_b} \in [0, 1]$
(2) $rin_j^{\mathbf{x}_b} \in [0, 1]$
(3) $rip_j^{\mathbf{x}_b} = 1$ *if* $rin_j^{\mathbf{x}_b} = 0$, $\forall \mathbf{x}_r \in \mathcal{I}(\mathbf{x}_b)$
(4) $rin_j^{\mathbf{x}_b} = 1$ *if* $rip_j^{\mathbf{x}_b} = 0$, $\forall \mathbf{x}_r \in \mathcal{I}(\mathbf{x}_b)$
(5) $rip_j^{\mathbf{x}_b} \neq 1$ *if* $rin_j^{\mathbf{x}_b} = 1$
(6) $rin_j^{\mathbf{x}_b} \neq 1$ *if* $rip_j^{\mathbf{x}_b} = 1$

Proof. (1) From Eq. (2), we deduce that $w_r^{\mathbf{x}_b} > w_s^{\mathbf{x}_b}$ if $Dis(\mathbf{x}_b, \mathbf{x}_r) < Dis(\mathbf{x}_b, \mathbf{x}_s)$, which means that for P instances with positive class on l_j, $rip_j^{\mathbf{x}_b}$ reaches maximum if all P instances are with the top P smallest $Dis(\mathbf{x}_b, \cdot)$, thus obtaining the maximum value of 1. Note that the value of $rip_j^{\mathbf{x}_b}$ decreases as more instances with negative class on l_j obtains smaller $Dis(\mathbf{x}_b, \cdot)$ and becomes the minimum if all P instances are with the top P largest $Dis(\mathbf{x}_b, \cdot)$. In the extreme case where all instances are with negative class, it reaches the minimum value of 0.

(2) This is similar to that of (1).

(3) Both $rin_j^{\mathbf{x}_b} = 0$ and $\forall \mathbf{x}_r \in \mathcal{I}(\mathbf{x}_b)$ imply that all instances in the neighborhood of $\mathbf{x}_b$ are with positive class on l_j, hence $rip_j^{\mathbf{x}_b} = 1$.

(4) This is similar to that of (3).

(5) $rin_j^{\mathbf{x}_b} = 1$ means all N instances are with the top N smallest $Dis(\mathbf{x}_b, \cdot)$, and $N > 1$ holds. Since the instance is either with the positive or negative class on l_j and $P = n - N$, there exists at least $\min(P, N)$ instances that are closer to $\mathbf{x}_b$ and with negative class simultaneously, which implies that $rip_j^{\mathbf{x}_b} < 1$. It is similar for that of $rip_j^{\mathbf{x}_b} = 1$.

(6) This is similar to that of (5). $\square$

The overall positive and negative tendencies can be estimated using the following triple tuples:

- Positive tendency: $\left(\mathcal{P}_j^{\mathbf{x}_b}, \mathbf{W}^{\mathbf{x}_b}, rip_j^{\mathbf{x}_b}\right)$
- Negative tendency: $\left(\mathcal{N}_j^{\mathbf{x}_b}, \mathbf{W}^{\mathbf{x}_b}, rin_j^{\mathbf{x}_b}\right)$

Consequently, the collection of the weighted score of the positive class on l_j is the assembly of all $\mathbf{x}_b$ and is defined as:

$$\mathbf{WSP}_j = \left(rip_j^{\mathbf{x}_b} \mathbf{W}^{\mathbf{x}_b\top} \mathcal{P}_j^{\mathbf{x}_b}\right)_{n\times 1} \tag{7}$$

Analogously, the weighted score of negative class on l_j is defined as:

$$\mathbf{WSN}_j = \left(rin_j^{\mathbf{x}_b} \mathbf{W}^{\mathbf{x}_b\top} \mathcal{N}_j^{\mathbf{x}_b}\right)_{n\times 1} \tag{8}$$

To objectively compare the local relative discrimination degrees (i.e., $\mathbf{WSP}_j - \mathbf{WSN}_j$), we simply consider the average distances within neighborhoods by defining the following augmented feature for l_j:

$$\mathbf{AF}_j = \frac{\mathbf{WSP}_j - \mathbf{WSN}_j}{Me(\mathbf{W}^{\mathbf{x}_b})} \tag{9}$$

where $Me(\mathbf{W}^{\mathbf{x}_b})$ denotes the median distance of the instances within the neighborhood of $\mathbf{x}_b$. It describes the compactness of k nearest neighborhood and can reduce the side-effect from outliers. The smaller the median value is, the more compact a k nearest neighborhood becomes. Therefore, the larger the component in $\mathbf{AF}_j$ is, the more reliable positive tendency of a k nearest neighborhood becomes, and vice versa.

3.4. Granular-based latent topic mining

Clustering is an effective strategy to characterize the topic structures. Instead of directly clustering completed on the original features, we employ k-means clustering on the concatenation of augmented features across the label space to generate the t topics. The advantages of generating topics in such a manner are three-fold.

1. From the viewpoint of concept cognition, people prefer to take the strategy of big concept priority. This means some abstraction procedures are required to construct the middle-level concepts like topics. In other words, the relationship between topics and the very detailed information (e.g., the original features) is rather weak.
2. The augmented features capture the distribution of neighborhood-based relative discrimination degrees. Instead, the relationship between original features and labels is ambiguous.
3. The clustering procedure will be accelerated, as the dimensionality of the augmented features is smaller than the original features.

Formally, we define $\mathbf{AF}^r \in \mathbb{R}^{n_r \times q}$ as the rth topic generated by the augmented features, where $\sum_r n_r = n$.

$$\left(\mathbf{AF}^1, \ldots, \mathbf{AF}^r, \ldots, \mathbf{AF}^t\right) = kmeans(\mathbf{AF}, t) \tag{10}$$

where $\mathbf{AF} = \left(\mathbf{AF}_1, \mathbf{AF}_2, \ldots, \mathbf{AF}_q\right)$ denotes the concatenation of augmented features on $l_1, l_2, \ldots, l_q$. $\mathbf{AF}^r = \left(\mathbf{AF}_1^r, \mathbf{AF}_2^r, \ldots, \mathbf{AF}_q^r\right)$ denotes the concatenation of augmented features on the rth topic. In this way, the instances with similar local relative discrimination degrees will be partitioned into the same topics, and the gaps between representation and classification are thus decreased.

3.5. Fine-grained label correlation learning

We generate label-specific features by learning an embedding matrix $\mathbf{V} \in \mathbb{R}^{m\times d}$ based on the partitions of the concatenation of both original and augmented features. This means the components of $\mathbf{V}$ are different if we consider different latent topics. Assuming that each label in a latent topic corresponds to a linear model as shown below:

$$f_j^r(\mathbf{x}) = \mathbf{u}_{jr}{}^\top \mathbf{V}_r^\top \mathbf{x} + b_{jr} \tag{11}$$

where $1 \leqslant j \leqslant q$, $\mathbf{u}_{jr}$ and b_{jr} correspond to the coefficient and bias in embedding feature space for the rth topic. To explore the contributions of features, MGOFA follows the granulation schema on data and knowledge by following the loss function below:

$$\min_{\mathbf{U}_r, \mathbf{V}_r, \mathbf{b}_r} \mathcal{L}\left(\mathbf{Y}_r, f\left(\mathbf{X}_{aug}^r, \mathbf{U}_r, \mathbf{V}_r, \mathbf{b}_r\right)\right) + \mathcal{R}_1(K) + \mathcal{R}_2(D) \tag{12}$$

where $\mathbf{U}_r \in \mathbb{R}^{d\times q}$, $\mathbf{V}_r \in \mathbb{R}^{m\times d}$, $\mathbf{b}_r \in \mathbb{R}^{q\times 1}$, and $\mathbf{1}_n \in \mathbb{R}^{n\times 1}$. Specifically, $\mathcal{L}\left(\mathbf{Y}_r, f\left(\mathbf{X}_{aug}^r, \mathbf{U}_r, \mathbf{V}_r, \mathbf{b}_r\right)\right)$ describes the structured loss of linear model, $\mathcal{R}_1(K)$ represents the correlation-based knowledge constraint. $\mathcal{R}_2(D)$ specifies regularization on model complexity. $\mathbf{X}_{aug}^r = \mathbf{X}^r \cup \mathbf{AF}^r$, where $\mathbf{X}^r$ stands for the original features of $\mathbf{X}$ w.r.t. rth topic. By considering

the topic-based second-order label correlation on embedded matrix, Eq. (12) can be expanded as:

$$\min_{\mathbf{U}_r,\mathbf{V}_r,\mathbf{b}_r} \frac{1}{2}\|\mathbf{X}^r_{aug}\mathbf{V}_r\mathbf{U}_r+\mathbf{1}_n\mathbf{b}_r^\top-\mathbf{Y}_r\|_2^2+\frac{\lambda_1}{2}tr\left(\mathbf{U}_r\mathbf{C}_r\mathbf{U}_r^\top\right)+\frac{\lambda_2}{2}\|\mathbf{V}_r\|_2^2+\lambda_3\|\mathbf{U}\|_1 \quad s.t.\ \mathbf{u}_{jr}^\top\mathbf{u}_{jr}=1, 1\leqslant j\leqslant q. \tag{13}$$

where $\frac{1}{2}\|\mathbf{X}^r_{aug}\mathbf{V}_r\mathbf{U}_r+\mathbf{1}_n\mathbf{b}_r^\top-\mathbf{Y}_r\|_2^2$ denotes the loss function of $\mathcal{L}\left(\mathbf{Y}_r,f\left(\mathbf{X}^r_{aug},\mathbf{U}_r,\mathbf{V}_r,\mathbf{b}_r\right)\right)$ on the rth topic. For the rth topic, $\mathcal{R}_1(K)=\frac{\lambda_1}{2}tr\left(\mathbf{U}_r\mathbf{C}_r\mathbf{U}_r^\top\right)$ measures the topic-based correlation between any two linear models w.r.t. l_a and l_b, which is expanded as $\sum_{a=1}^q\sum_{b=1}^q c_{ab}\mathbf{u}_{ar}^\top\mathbf{u}_{br}$, $\mathcal{R}_2(D)=\frac{\lambda_2}{2}\|\mathbf{V}_r\|_2^2+\lambda_3\|\mathbf{U}\|_1$ constraints the model complexity on embedding matrix and weight sparsity. $\mathbf{C}_r=\left[c^r_{ab}\right]_{q\times q}\in\mathbb{R}^{q\times q}$, where $c_{ab}=-\sum_{i=1}^n y_{ia}y_{ib}$ denotes the negation of the number of instances which is associated with l_a and l_b simultaneously in rth topic. By imposing this constraint, we enforce the weights for l_a and l_b in rth topic should be similar if l_a and l_b for instances in rth topic are strongly correlated, and vice versa.

The objective function in Eq. (13) can be solved via alternating optimization as below.

(1) **Update $\mathbf{U}_r$, with $\mathbf{V}_r$ and $\mathbf{b}_r$ fixed:**

Eq. (13) reduces to:

$$\min_{\mathbf{U}_r} f\left(\mathbf{U}_r\right)+\lambda_3\left\|\mathbf{U}_r\right\|_1 \quad s.t.\ \mathbf{u}_{jr}^\top\mathbf{u}_{jr}=1,\ 1\leqslant j\leqslant q \tag{14}$$

where $f\left(\mathbf{U}_r\right)=\frac{1}{2}\left\|\mathbf{X}^r_{aug}\mathbf{V}_r\mathbf{U}_r+\mathbf{1}_n\mathbf{b}^\top-\mathbf{Y}_r\right\|_2^2+\frac{\lambda_1}{2}tr\left(\mathbf{U}_r\mathbf{C}_r\mathbf{U}_r^\top\right)$ corresponds to the smoothing part of convex objective function in Eq. (14). By invoking proximal gradient descent, we deduce the gradient of the objective w.r.t. $\mathbf{U}_r$ on $f\left(\mathbf{U}_r\right)$ as:

$$\nabla_{\mathbf{U}_r}f\left(\mathbf{U}_r\right)=\left(\mathbf{X}^r_{aug}\mathbf{V}_r\right)^\top\left(\mathbf{X}^r_{aug}\mathbf{V}_r\mathbf{U}_r+\mathbf{1}_n\mathbf{b}_r^\top-\mathbf{Y}_r\right)+\lambda_1\mathbf{U}_r\mathbf{C}_r \tag{15}$$

Furthermore, $f\left(\mathbf{U}_r\right)$ satisfies the L-Lipschitz condition as:

$$\left\|\nabla_{\mathbf{U}_r}f\left(\mathbf{U}_r^1\right)-\nabla_{\mathbf{U}}f\left(\mathbf{U}_r^2\right)\right\|_2 \leqslant\left(\left\|\mathbf{X}^r_{aug}\mathbf{V}_r\right\|_2^2+\lambda_1\left\|\mathbf{Y}_r\right\|_2^2\right)\cdot\left\|\mathbf{U}_r^1-\mathbf{U}_r^2\right\| \tag{16}$$

with Lipschitz constant $L=\left\|\mathbf{X}^r_{aug}\mathbf{V}_r\right\|_2^2+\lambda_1\left\|\mathbf{Y}_r\right\|_2^2$. Therefore, the weight matrix $\mathbf{U}_r$ at the round $iter+1$ can be iteratively updated based on the previous round as:

$$\mathbf{U}^{(iter+1)}_{r,ij}=\frac{\mathbf{S}^{(iter)}_{r,ij}}{\left\|\mathbf{s}_{r,j}\right\|_2} \tag{17}$$

where $\mathbf{S}^{(iter)}_{r,ij}=sgn\left(\mathbf{Z}^{(iter)}_{r,ij}\right)\max\left(\left|\mathbf{Z}^{(iter)}_{r,ij}\right|-\frac{\lambda_3}{L},0\right)$ and $\mathbf{Z}^{(iter)}_r=\mathbf{U}^{(iter)}_r-\frac{1}{L}\nabla_{\mathbf{U}}f\left(\mathbf{U}^{(iter)}\right)$. $\mathbf{s}_{r,j}=\left[\mathbf{S}^{(iter)}_{r,1j},\ldots,\mathbf{S}^{(iter)}_{r,dj}\right]^\top$.

Meanwhile, $\mathbf{C}_r$ is renewed by:

$$\mathbf{C}_r=\mathbf{X}^r_{aug}{}^\top\left(\mathbf{1}_n\mathbf{b}_r^\top-\mathbf{Y}_r\right)\mathbf{U}_r^\top \tag{18}$$

(2) **Update $\mathbf{V}_r$ and $\mathbf{b}_r$, with $\mathbf{U}_r$ fixed:**

Eq. (13) reduces to:

$$\min_{\mathbf{V}_r,\mathbf{b}_r}\frac{1}{2}\left\|\mathbf{X}^r_{aug}\mathbf{V}_r\mathbf{U}_r+\mathbf{1}_n\mathbf{b}_r^\top-\mathbf{Y}_r\right\|_2^2+\frac{\lambda_2}{2}\left\|\mathbf{V}_r\right\|_2^2 \tag{19}$$

The solving on $\mathbf{V}_r$ satisfies the following condition:

$$\mathbf{X}^r_{aug}{}^\top\mathbf{X}^r_{aug}\mathbf{V}_r\mathbf{U}_r\mathbf{U}_r^\top+\mathbf{X}^r_{aug}{}^\top\left(\mathbf{1}_n\mathbf{b}_r^\top-\mathbf{Y}_r\right)\mathbf{U}_r^\top+\lambda_2\mathbf{V}_r=\mathbf{0} \tag{20}$$

Let $\mathbf{E}_r=\mathbf{X}^r_{aug}{}^\top\left(\mathbf{1}_n\mathbf{b}_r^\top-\mathbf{Y}_r\right)\mathbf{U}_r^\top$, $\mathbf{A}_r=\mathbf{X}^r_{aug}{}^\top\mathbf{X}^r_{aug}$, and $\mathbf{B}_r=\mathbf{U}_r\mathbf{U}_r^\top$, the symmetric matrix $\mathbf{A}_r$ can be factorized as $\mathbf{P}_r\boldsymbol{\Lambda}_r\mathbf{P}_r^\top$, where $\mathbf{P}_r$ is an orthonormal matrix whose columns are the matrix of $\mathbf{A}_r$ and $\boldsymbol{\Lambda}_r$ is a diagonal matrix whose diagonal elements are the eigenvalues of matrix $\mathbf{A}_r$. The factorization on $\mathbf{B}_r$ is similar and can be denoted by $\mathbf{Q}_r\boldsymbol{\Gamma}_r\mathbf{Q}_r^\top$, where $\mathbf{Q}_r$ is an orthonormal matrix and $\boldsymbol{\Gamma}_r$ is a diagonal matrix with the elements as eigenvalues of $\mathbf{Q}_r$. Consequently, Eq. (20) can be rewritten as:

$$\mathbf{P}_r\boldsymbol{\Lambda}_r\mathbf{P}_r^\top\mathbf{V}_r\mathbf{Q}_r\boldsymbol{\Gamma}_r\mathbf{Q}_r^\top+\mathbf{E}_r+\lambda_2\mathbf{V}_r=\mathbf{0} \tag{21}$$

By multiplying $\mathbf{P}_r^\top$ and $\mathbf{Q}_r$ to the left side and right side of Eq. (21), we have:

$$\boldsymbol{\Lambda}_r\mathbf{P}_r^\top\mathbf{V}_r\mathbf{Q}_r\boldsymbol{\Gamma}_r+\mathbf{P}_r^\top\mathbf{E}_r\mathbf{Q}_r+\lambda_2\mathbf{P}_r^\top\mathbf{V}_r\mathbf{Q}_r=\mathbf{0} \tag{22}$$

The closed-form solution of $\mathbf{V}_r$ is thus denoted as:

$$\mathbf{V}_r=\mathbf{P}_r\left(\left(-\mathbf{P}_r^\top\mathbf{E}_r\mathbf{Q}_r\right)\oslash\left(\boldsymbol{\Lambda}_r\mathbf{1}_n\mathbf{1}_d^\top\boldsymbol{\Gamma}_r+\lambda_2\mathbf{1}_n\mathbf{1}_d^\top\right)\right)\mathbf{Q}_r^\top \tag{23}$$

where $\oslash$ represents the Hadamard division operator. The closed form solution of $\mathbf{b}_r$ is denoted as:

$$\mathbf{b}_r=-\frac{1}{n}\left(\mathbf{X}^r_{aug}\mathbf{V}_r\mathbf{U}_r-\mathbf{Y}_r\right)^\top\mathbf{1}_n \tag{24}$$

3.6. Bias analysis with feature augmentation

A rigorous theoretical analysis on whether feature augmentation incurs additional bias is essential. Ensuring that any performance improvements from feature augmentation are not accompanied by increased bias is crucial for maintaining the reliability and robustness of the proposed MGOFA. Theorem 1 is formally given to address this concern.

Theorem 1. *Let* $\mathbf{X}$ *and* $\mathbf{AF}$ *denote the original and augmented features, where augmented features are generated based on the approach covered in Section 3.3.2,* $\mathbf{X}_{aug}=\mathbf{X}\cup\mathbf{AF}$. *Suppose* $Bias_o(\mathbf{X})$, $Bias_c\left(\mathbf{X}_{aug}\right)$, $Bias^r_m\left(\mathbf{X}^r_{aug}\right)$ *denote the bias of learning classifier based on original, concatenating features, and multi-granularity features, where the concatenating features and multi-granularity features is based on studies in Sections 3.3.2 and 3.5, respectively. Then* $E\left(Bias^r_m\left(\mathbf{X}^r_{aug}\right)\right)<E\left(Bias_c\left(\mathbf{X}_{aug}\right)\right)<E\left(Bias_o(\mathbf{X})\right)$ *is held, where* $E(\cdot)$ *represents the expectation operator.*

Proof. In the terminology, bias stands for the difference between the expected output and the ground-truth labels. For the term $Bias_o(\mathbf{X})$, we have:

$$Bias_o(\mathbf{X})=E\left(f_o(\mathbf{X})\right)-\mathbf{Y}$$

where $f_o(\mathbf{X})$ denotes the output with the original features being treated as input, while $\mathbf{Y}$ denotes the ground-truth labels.

Similarly, we have $Bias_c\left(\mathbf{X}_{aug}\right)$ and $Bias^r_m\left(\mathbf{X}^r_{aug}\right)$ expressed as:

$$Bias_c\left(\mathbf{X}_{aug}\right)=E\left(f_c\left(\mathbf{X}_{aug}\right)\right)-\mathbf{Y}$$

$$Bias^r_m\left(\mathbf{X}^r_{aug}\right)=E\left(f_m\left(\mathbf{X}^r_{aug}\right)\right)-\mathbf{Y}$$

where $f_c\left(\mathbf{X}_{aug}\right)$ denotes the output with the concatenation of original and augmented features treated as input, while $f_m\left(\mathbf{X}^r_{aug}\right)$ denotes the output with multi-granularity features as input. In what follows, we demonstrate the inequality assertion in two steps:

(1) Firstly, we prove that $E\left(Bias_c\left(\mathbf{X}_{aug}\right)\right)<E\left(Bias_o(\mathbf{X})\right)$:

As suggested by Eq. (9), each element of $\mathbf{AF}$ (i.e., $\mathbf{AF}_i$) strengthens the relevancy between augmented feature and corresponding labels while maintaining the neighborhood-based similarity of raw data, which means

$$H\left(l_i \left| \mathbf{X}_{aug} \right.\right) < H\left(l_i \left| \mathbf{X} \right.\right)$$

where $H\left(l_i \left| \cdot \right.\right)$ denotes the conditional entropy of determining l_i. Therefore, the concatenation of original and augmented features offers a finer discrimination on label association. In other words, under the sparse feature coefficients constraint, the sparse coefficient combinations generated by $f_o(\mathbf{X})$ is a subset on that of $f_c\left(\mathbf{X}_{aug}\right)$. The functionality of $\mathbf{u}_c$ can also be observed from Eq. (11) as:

$$\mathbf{u}_c = \left(\mathbf{u}_o, \mathbf{u}_a\right)$$

where $\left(\mathbf{u}_o, \mathbf{0}\right)$ is a special case of $\mathbf{u}_c$, which means that all augmented features do not contribute to the classification. Given that the conditional entropy is decreasing, we have:

$$E\left(Bias_o\left(\mathbf{X}\right)\right) = E\left(Bias_c\left(\mathbf{X}_{aug}\right)\right) + \Delta_{oc}$$

where $\Delta_{oc} > 0$ represents the improvement of feature augmentation on bias estimation.

(2) Secondly, we prove that $E\left(Bias_m^r\left(\mathbf{X}_{aug}^r\right)\right) < E\left(Bias_c\left(\mathbf{X}_{aug}\right)\right)$:

The latent topics deduced by the similarity of augmented features generate a partition of $\mathbf{X}$. In other words, there is a group of $\mathbf{u}_c$, where each of them is different in terms of sparse combinations. Suppose two arbitrary labels l_a and l_b shows salient label correlation w.r.t. rth topic. Then for the concatenation of original and augmented features on rth topic, we have:

$$H\left(\mathbf{C}_r \left| \mathbf{X}_{aug}^r \right.\right) < H\left(\mathbf{C}_r \left| \mathbf{X}_{aug} \right.\right)$$

where $\mathbf{C}_r$ denotes the correlation matrix in rth topic, and $H\left(\mathbf{C}_r \left| \cdot \right.\right)$ measures the conditional entropy of fitting the latent label correlation in $\mathbf{C}_r$.

The reduction in conditional entropy means that given the feature representation of the rth topic (i.e., $\mathbf{X}_{aug}^r$), the uncertainty of the corresponding local label correlation (i.e., $\mathbf{C}_r$) is reduced. As the robust estimation on label correlation is conducive to deducing accurate classification, the capability of classifying l_i becomes stronger, which implies the decrease of $H\left(l_i \left| \mathbf{C}_r \right.\right)$. Given that $\mathbf{C}_r$ is deduced by $\mathbf{X}_{aug}^r$, it implies:

$$H\left(l_i \left| \mathbf{X}_{aug}^r \right.\right) < H\left(l_i \left| \mathbf{X}_{aug} \right.\right)$$

Thus, we have:

$$E\left(Bias_c\left(\mathbf{X}_{aug}\right)\right) = E\left(Bias_m^r\left(\mathbf{X}_{aug}^r\right)\right) + \Delta_{cm}$$

where $\Delta_{cm} > 0$ represents the bias improvement of finding latent topics based on feature augmentation.

Combining (1) and (2), the $E\left(Bias_m^r\left(\mathbf{X}_{aug}^r\right)\right) < E\left(Bias_c\left(\mathbf{X}_{aug}\right)\right) < E\left(Bias_o\left(\mathbf{X}\right)\right)$ is held. □

Theorem 1 demonstrates that the feature augmentation is conducive to reducing the bias in constructing the proposed MGOFA. The values of Δ_{oc} and Δ_{cm} are pertinent to the dataset characteristics, and will be measured experimentally in the ablation study (see Section 4.7).

3.7. Complexity analysis

Algorithm 1 summarizes the major procedures of the MGOFA. The training procedures of MGOFA include "Granular-based Feature Augmentation" (Step 1–Step 10), "Granular-based Latent Topic Mining" (Step 11), and "Fine-grained Label Correlation Learning" (Step 12–Step 25). Step 26 corresponds to the testing procedure.

The complexity of MGOFA is analyzed as follows. The complexity of "Granular-based Feature Augmentation" is $O\left(n^2m\right)+O\left(nkq\right)$, which corresponds to the instance-based neighborhood generation and relative class tendency estimation, respectively. The complexity of "Granular-based Latent Topic Mining" is $O\left(n^2q\right)$. The complexity of "Fine-grained Label Correlation Learning" in a single iteration is $O\left(n_r^2r^2m^2d^2q^2\right)$, where n_r denotes the instance count partitioned into the rth topic. Given that both $d < n_r$ and $q < m$ hold in most cases, the overall complexity of MGOFA is $O\left(n_r^2r^2m^2d^2q^2\right)$.

Algorithm 1: Multi-Granularity cOrrelation-based Feature Augmentation(MGOFA)

Input: Training set $\mathbf{X}$, ground-truth labels $\mathbf{Y}$, Nearest neighborhood count k, distance balance factor μ, topic number t, embedding dimension d, regularization parameters λ_1, λ_2, and λ_3, unseen instances $\mathbf{X}^*$

Output: Predicted labels $\mathbf{Y}^*$

1 **for** $b = 1$ *to* n **do**
2 Find the k nearest neighborhood w.r.t. $\mathbf{x}_b$ based on Eq. (1).
3 Compute the instance weight $\mathbf{W}^{\mathbf{x}_b}$ w.r.t. $\mathbf{x}_b$ based on Eq. (2).
4 **for** $j = 1$ *to* q **do**
5 Find the positive and negative class (i.e., $\mathcal{P}_j^{\mathbf{x}_b}$, $\mathcal{N}_j^{\mathbf{x}_b}$) based on Eqs. (3) and (4).
6 Compute the relative influential ratio of positive and negative class (i.e., $\mathbf{rip}_j$ and $\mathbf{rin}_j$) based on Eqs. (5) and (6).
7 Compute weighted score of positive and negative class (i.e., $\mathbf{WSP}_j$ and $\mathbf{WSN}_j$) based on Eqs. (7) and (8).
8 Generate the jth augmented feature $\mathbf{AF}_j$ based on Eq. (9).
9 **end**
10 **end**
11 Generate t topic-based representation $\{\mathbf{X}_{aug}^1, \mathbf{X}_{aug}^2, \ldots, \mathbf{X}_{aug}^r, \ldots, \mathbf{X}_{aug}^t\}$ based on Eq. (10).
12 **for** $r = 1$ *to* t **do**
13 Factorize the ${\mathbf{X}_{aug}^r}^\top \mathbf{X}_{aug}^r$ into $\mathbf{P}_r\boldsymbol{\Lambda}_r\mathbf{P}_r^\top$
14 **repeat**
15 Randomly initialize $\mathbf{U}_r$, $\mathbf{b}_r$, and $\mathbf{V}_r$ s.t. $\mathbf{U}_r\mathbf{U}_r^\top = \mathbf{I}$, $iter \leftarrow 0$.
16 **repeat**
17 Compute $\bigtriangledown_{\mathbf{U}_r} f\left(\mathbf{U}_r\right)$ based on Eq. (15) and update $\mathbf{U}_r^{(iter+1)}$ based on Eq. (17).
18 $iter \leftarrow iter + 1$.
19 **until** *convergence*;
20 Factorize $\mathbf{U}_r\mathbf{U}_r^\top$ into $\mathbf{Q}_r\boldsymbol{\Gamma}_r\mathbf{Q}_r^\top$, where $\mathbf{U}_r \leftarrow \mathbf{U}_r^{(iter)}$.
21 Update $\mathbf{C}_r$ by Eq. (18).
22 Update $\mathbf{V}_r$ by Eq. (23).
23 Update $\mathbf{b}_r$ by Eq. (24).
24 **until** *convergence*;
25 **end**
26 Return $\mathbf{Y}^* = \left\{\omega_j \left| f_j^r(\mathbf{x}) > 0.5, 1 \leqslant j \leqslant q \right.\right\}$

4. Experiments

4.1. Datasets

Table 3 enumerates characteristics of ten datasets, including the instance count $|S|$, feature dimensionality $dim(S)$, label count $L(S)$, and the cardinality of average associated labels per instance $LCard(S)$. They all come from Lamda[1] repository.

4.2. Experimental settings

To evaluate the effectiveness of MGOFA, we adopt five widely used metrics [48], including *Hamming Loss*, *Ranking Loss*, *One Error*, *Average Precision* and *Macro-averaging AUC*. Except for the last two metrics that signify a better performance for larger values, the remaining embrace smaller values for better performance.

We compare MGOFA against eight state-of-the-art multi-label algorithms for performance evaluations to examine whether MGOFA

[1] https://www.lamda.nju.edu.cn/code_MDDM.ashx.

Table 3
Dataset characteristics.

Data set	$\|S\|$	$dim(S)$	$L(S)$	$LCard(S)$
Art	5000	462	26	1.64
Business	5000	438	30	1.59
Computers	5000	681	33	1.51
Education	5000	550	33	1.46
Entertainment	5000	640	21	1.42
Health	5000	612	32	1.66
Recreation	5000	606	22	1.42
Reference	5000	793	33	1.17
Science	5000	743	40	1.45
Social	5000	1047	39	1.28

achieves better performance than solutions that are (1) with optimal components via feature selection and (2) with stronger label correlation assumption on such optimal components. The configurations for these algorithms take the recommended values via five-fold cross-validation.

- WRAP [28][23]: A label-specific multi-label classification that takes a wrapped approach w.r.t. each label by considering global label correlation. [parameter configurations: grid search for $\lambda_1, \lambda_2 \in \{0, 1, \ldots, 10\}$. $\lambda_3 \in \{0.1, 1\}$ and $\alpha = 0.9$].
- HOMI [43][4]: A high-order label correlation learning with self-representation and local geometric structure on global label correlation. [parameter configurations: grid search for $\beta, \lambda, \gamma \in \{10^{-5}, 10^{-4}, \ldots, 1\}$ and $s \in \{5, 10\}$].
- SLOFS [44][5]: A shared latent sublabel structure with global label correlation. [parameter configurations: grid search for $\alpha, \lambda_1, \lambda_2, \beta, \delta \in \{0.01, 0.1, 0.3, 0.5, 0.7, 0.9, 1\}$]
- MDFS [15][6]: An embedded feature selection strategy with manifold regularization to exploit discriminative features on both global and local label correlation. [parameter configurations: $\alpha = 1$, grid search for $\beta, \gamma \in \{10^{-3}, 10^{-2}, \ldots, 10^{3}\}$].
- TIFS [16][7]: A latent topic-based instance and feature selection with global and local label correlation simultaneously. [parameter configurations: $\lambda = 0.5, k = 10, s = 50$, grid search for $\tau, \delta \in \{10^{-4}, 10^{-3}, \ldots, 10^{-1}\}$].
- RLFSCL [45][8]: A low-rank feature and label representation learning approach with global label correlation. [parameter configurations: $\rho = 1.1$, grid search for $\lambda_1, \lambda_2 \in \{10^{-3}, 10^{-2}, \ldots, 10^{3}\}$, $\mu \in \{1, 10^{1}, \ldots, 10^{6}\}$].
- GLFS [46][9]: A group-preserving label-specific feature selection learning for global and local label correlation learning. [parameter configurations: grid search for $\alpha, \lambda \in \{0, 0.2, \ldots, 1\}$, $K = 5, M = 16$, $\beta, \gamma \in \{10^{-3}, 10^{-2}, \ldots, 10^{3}\}$].
- MC-GM [47][10]: A group-specific feature selection strategy with label-specific group selection by combining global and local label correlation. [parameter configurations: grid search for $\lambda, \beta \in \{10^{-4}, 10^{-2}, \ldots, 1\}$, $\delta, \alpha \in \{10^{-2}, 10^{-1}, \ldots, 10^{2}\}$, $s = 50$].

The configurations of MGOFA are as follows: The balance factor μ is searched in $[0, 1]$ at a step of 0.1. We stipulate $k = 10$ for neighborhood construction and $t = 3$ as topic count. The d dimensionality in embedded matrix $\mathbf{V}$ are defined as $d = \lceil \beta \min(m, q) \rceil$, where $\beta = 0.9$. We take a grid search manner for trade-off $\lambda_1, \lambda_2 \in \{1, 2, \ldots, 10\}$. The trade-off λ_3 is fixed as $\lambda_3 = 1$.

For fair comparisons, all experiments are executed via five-fold cross-validation on a desktop with Intel i7-10700CPU (2.90 GHz) and 32 GB RAM. The MGOFA is implemented via Matlab R2017b.

4.3. Results

Table 4 shows the detailed classification performance ranking information over ten benchmarks and eight comparing methods. From the metric view, MGOFA ranks first at 80% cases $\left(\frac{4}{5}\right)$, ranks second at 20% cases $\left(\frac{1}{5}\right)$. From the dataset view, MGOFA ranks first at 70% $\left(\frac{35}{50}\right)$ cases, ranks second at 16%$\left(\frac{8}{50}\right)$ cases, ranks third at 12%$\left(\frac{6}{50}\right)$ cases.

We employ the Friedman test [49] to examine whether the statistical difference in relative performance holds for all metrics. Let N, T, and r_i^j denote the comparison approaches count, the dataset count, and the rank of the jth algorithm on the ith dataset, respectively. Given the average rank (i.e., $R_j = \frac{1}{T}\sum_{i=1}^{T} r_i^j$) information induced in Table 4, Friedman statistic F_F follows the F-distribution under the null hypothesis that all algorithms are statistically indistinguishable, with $N - 1$ numerator degrees of freedom and $(N-1)(T-1)$ denominator degrees of freedom.

$$F_F = \frac{(T-1)\chi_F^2}{T(N-1) - \chi_F^2} \tag{25}$$

where $\chi_F^2 = \frac{12T}{N(N+1)}\left[\sum_{j=1}^{N} R_j^2 - \frac{N(N+1)^2}{4}\right]$. Table 5 summarizes the results for all considered five metrics, where $N = 9$ and $T = 10$. Given that critical value $\alpha = 0.05$, the null hypothesis of statistically indistinguishable performance among all considered algorithms is rejected for all metrics.

To further examine whether MGOFA is significantly superior over the remaining algorithms on different metrics, we employ Holm's procedure [49] as the post-hoc test by regarding MGOFA as the control algorithm (denoted as $\mathcal{A}_1$). For the remaining $T - 1$ comparing algorithms (denoted as $\mathcal{A}_j$, where $2 \leqslant j \leqslant N$), the one obtaining the $j-1$th largest average rank over all datasets is denoted as $\mathcal{A}_j$. Then we have the test statistic for comparing $\mathcal{A}_1$ (i.e., MGOFA) and $\mathcal{A}_j$ in Eq. (26).

$$z_j = (R_1 - R_j)/\sqrt{\frac{N(N+1)}{6T}} \quad (2 \leqslant j \leqslant N) \tag{26}$$

In practice, we stipulate p_j as the p-value of z_j under the normal distribution. Given the significance level $\alpha = 0.05$, the Holm's procedure works by sequentially examining whether $p_j < \alpha/(N - j + 1)$holds in ascending order of j. Typically, the Holm's procedure continues until the first j (denoted as j^*) which $p_j \geqslant \alpha/(N - j + 1)$ holds.[11] Consequently, MGOFA is statistically superior over algorithms with the ranking of $\mathcal{A}_j$, where $j \in \{2, \ldots, j^* - 1\}$.

Table 6 enumerates the results of Holm's procedure. We infer that MGOFA achieves the most statistically superior at metric *Hamming Loss*. Additionally, it achieves statistically superior performance over TIFS, SLOFS, on all five metrics, GLFS, MDFS on all metrics except for *Ranking Loss*, MC-GM on *Hamming Loss* and *Ranking Loss*.

4.4. Convergence analysis

Fig. 5 illustrates the convergence of objective functions w.r.t. different topics with the parameters $(\beta, k, \mu, t, \lambda_1, \lambda_2, \lambda_3) = (0.9, 10, 0.4, 3, 5, 5, 1)$, on datasets *Art*, *Business*, and *Computers*, respectively. The convergence condition is either the variations of the objective function in two consecutive rounds are smaller than 10^{-4} or the iteration count reaches 1000. From Fig. 5, we can observe that the objective function of MGOFA decreases rapidly in a limited number of iterations for all topics. The maximal round of reaching convergence is smaller than 400, which implies the solution is optimal.

[2] https://palm.seu.edu.cn/zhangml/files/WRAP.rar.
[3] For fair comparisons, we use the linear version instead of the kernel version.
[4] https://github.com/Chongjie-Si/HOMI.
[5] https://github.com/zhongjingyu1/SLOFS.
[6] https://github.com/jiazhang-ml/MDFS.
[7] https://github.com/JianghongMA/TIFS.
[8] https://github.com/JingChuanTang/RLFSCL.
[9] https://github.com/jiazhang-ml/GLFS.
[10] https://github.com/JianghongMA/MC-GM.
[11] j^* is set to be $N + 1$ if $p_j < \alpha/(N - j + 1)$ holds $\forall j$.

Table 4
Comparisons (mean ± std) on five metrics (↓: the smaller the better, ↑: the larger the better).

Dataset	*Hamming loss* ↓								
	MGOFA	WRAP	HOMI	SLOFS	MDFS	TIFS	RLFSCL	GLFS	MC-GM
Art	**0.054 ± 0.001**	**0.054 ± 0.001**	0.060 ± 0.001	0.063 ± 0.001	0.063 ± 0.001	0.124 ± 0.020	0.057 ± 0.001	0.063 ± 0.001	0.086 ± 0.002
Business	**0.025 ± 0.000**	**0.025 ± 0.001**	0.026 ± 0.001	0.029 ± 0.001	0.028 ± 0.001	0.044 ± 0.006	0.026 ± 0.000	0.028 ± 0.002	0.032 ± 0.002
Computers	**0.034 ± 0.001**	**0.034 ± 0.001**	0.035 ± 0.001	0.040 ± 0.001	0.038 ± 0.001	0.074 ± 0.008	0.035 ± 0.001	0.039 ± 0.001	0.063 ± 0.005
Education	**0.037 ± 0.001**	0.038 ± 0.001	0.043 ± 0.000	0.044 ± 0.001	0.044 ± 0.001	0.072 ± 0.007	0.039 ± 0.001	0.044 ± 0.000	0.060 ± 0.001
Entertainment	**0.052 ± 0.002**	**0.052 ± 0.001**	0.063 ± 0.002	0.068 ± 0.001	0.067 ± 0.001	0.112 ± 0.008	0.055 ± 0.001	0.067 ± 0.001	0.087 ± 0.004
Health	0.035 ± 0.001	0.035 ± 0.001	0.036 ± 0.001	0.050 ± 0.001	0.036 ± 0.001	0.044 ± 0.001	**0.032 ± 0.001**	0.047 ± 0.001	0.048 ± 0.002
Recreation	**0.054 ± 0.001**	**0.054 ± 0.001**	0.062 ± 0.001	0.065 ± 0.001	0.065 ± 0.001	0.124 ± 0.014	0.056 ± 0.001	0.065 ± 0.001	0.098 ± 0.007
Reference	**0.026 ± 0.001**	**0.026 ± 0.001**	0.035 ± 0.000	0.031 ± 0.001	0.029 ± 0.001	0.051 ± 0.004	0.028 ± 0.001	0.035 ± 0.000	0.050 ± 0.002
Science	**0.032 ± 0.001**	**0.032 ± 0.001**	0.035 ± 0.000	0.036 ± 0.001	0.036 ± 0.000	0.061 ± 0.004	0.034 ± 0.001	0.036 ± 0.000	0.057 ± 0.001
Social	**0.021 ± 0.001**	**0.021 ± 0.001**	0.025 ± 0.001	0.028 ± 0.001	0.024 ± 0.001	0.042 ± 0.004	0.022 ± 0.001	0.026 ± 0.001	0.033 ± 0.003
Dataset	*Ranking loss* ↓								
	MGOFA	WRAP	HOMI	SLOFS	MDFS	TIFS	RLFSCL	GLFS	MC-GM
Art	0.135 ± 0.005	0.133 ± 0.006	**0.124 ± 0.006**	0.176 ± 0.004	0.172 ± 0.002	0.189 ± 0.018	0.135 ± 0.006	0.174 ± 0.003	0.156 ± 0.006
Business	**0.040 ± 0.003**	**0.040 ± 0.002**	**0.040 ± 0.003**	0.048 ± 0.004	0.048 ± 0.002	0.050 ± 0.009	0.061 ± 0.006	0.047 ± 0.004	0.054 ± 0.003
Computers	**0.092 ± 0.009**	0.094 ± 0.008	0.102 ± 0.008	0.096 ± 0.005	0.095 ± 0.001	0.136 ± 0.013	0.097 ± 0.004	0.094 ± 0.004	0.127 ± 0.010
Education	0.106 ± 0.008	0.107 ± 0.007	**0.093 ± 0.003**	0.108 ± 0.004	0.106 ± 0.003	0.128 ± 0.011	0.097 ± 0.006	0.108 ± 0.003	0.127 ± 0.010
Entertainment	0.113 ± 0.007	0.114 ± 0.007	**0.100 ± 0.005**	0.155 ± 0.004	0.137 ± 0.003	0.152 ± 0.005	0.115 ± 0.002	0.149 ± 0.002	0.128 ± 0.050
Health	**0.062 ± 0.004**	0.063 ± 0.005	**0.062 ± 0.003**	0.082 ± 0.003	0.070 ± 0.006	0.056 ± 0.003	0.075 ± 0.006	0.073 ± 0.006	0.079 ± 0.006
Recreation	0.149 ± 0.004	0.146 ± 0.007	**0.136 ± 0.009**	0.213 ± 0.005	0.197 ± 0.003	0.212 ± 0.013	0.149 ± 0.006	0.205 ± 0.005	0.162 ± 0.008
Reference	**0.093 ± 0.006**	0.094 ± 0.003	0.189 ± 0.007	0.110 ± 0.006	0.112 ± 0.004	0.107 ± 0.007	0.095 ± 0.008	0.110 ± 0.003	0.110 ± 0.010
Science	0.134 ± 0.010	0.131 ± 0.007	**0.119 ± 0.006**	0.153 ± 0.002	0.149 ± 0.005	0.152 ± 0.009	0.133 ± 0.006	0.149 ± 0.002	0.138 ± 0.006
Social	0.082 ± 0.004	0.082 ± 0.003	**0.065 ± 0.003**	0.078 ± 0.004	0.078 ± 0.002	0.091 ± 0.004	0.084 ± 0.007	0.078 ± 0.005	0.095 ± 0.008
Dataset	*One error* ↓								
	MGOFA	WRAP	HOMI	SLOFS	MDFS	TIFS	RLFSCL	GLFS	MC-GM
Art	**0.460 ± 0.010**	0.462 ± 0.018	0.492 ± 0.006	0.749 ± 0.009	0.730 ± 0.012	0.689 ± 0.066	0.524 ± 0.020	0.738 ± 0.010	0.471 ± 0.014
Business	**0.111 ± 0.004**	0.113 ± 0.008	0.114 ± 0.016	0.135 ± 0.011	0.135 ± 0.004	0.293 ± 0.054	0.115 ± 0.007	0.135 ± 0.008	0.127 ± 0.018
Computers	0.353 ± 0.018	**0.350 ± 0.009**	0.371 ± 0.014	0.476 ± 0.020	0.476 ± 0.012	0.616 ± 0.041	0.383 ± 0.010	0.476 ± 0.011	0.368 ± 0.014
Education	**0.468 ± 0.018**	0.471 ± 0.006	0.507 ± 0.015	0.685 ± 0.007	0.685 ± 0.008	0.689 ± 0.049	0.519 ± 0.015	0.685 ± 0.017	0.477 ± 0.009
Entertainment	**0.399 ± 0.017**	0.403 ± 0.015	0.436 ± 0.022	0.715 ± 0.009	0.650 ± 0.009	0.608 ± 0.031	0.459 ± 0.010	0.700 ± 0.025	0.404 ± 0.007
Health	0.273 ± 0.010	0.270 ± 0.013	0.283 ± 0.011	0.493 ± 0.016	0.289 ± 0.013	0.364 ± 0.020	**0.254 ± 0.008**	0.448 ± 0.021	0.324 ± 0.024
Recreation	0.461 ± 0.006	**0.459 ± 0.021**	0.488 ± 0.022	0.805 ± 0.012	0.750 ± 0.024	0.691 ± 0.027	0.521 ± 0.017	0.799 ± 0.011	0.469 ± 0.018
Reference	**0.373 ± 0.010**	0.381 ± 0.009	0.535 ± 0.005	0.535 ± 0.015	0.535 ± 0.009	0.621 ± 0.045	0.423 ± 0.007	0.535 ± 0.016	0.384 ± 0.020
Science	**0.499 ± 0.019**	0.501 ± 0.029	0.524 ± 0.016	0.728 ± 0.015	0.682 ± 0.027	0.706 ± 0.037	0.565 ± 0.013	0.699 ± 0.011	0.502 ± 0.015
Social	0.294 ± 0.015	**0.289 ± 0.016**	0.341 ± 0.007	0.446 ± 0.014	0.404 ± 0.004	0.554 ± 0.047	0.325 ± 0.015	0.431 ± 0.017	0.294 ± 0.016
Dataset	*Average precision* ↑								
	MGOFA	WRAP	HOMI	SLOFS	MDFS	TIFS	RLFSCL	GLFS	MC-GM
Art	**0.583 ± 0.008**	**0.583 ± 0.009**	0.573 ± 0.013	0.419 ± 0.005	0.429 ± 0.006	0.441 ± 0.047	0.555 ± 0.013	0.424 ± 0.004	0.572 ± 0.011
Business	**0.856 ± 0.006**	0.855 ± 0.004	0.854 ± 0.008	0.826 ± 0.010	0.827 ± 0.004	0.777 ± 0.033	0.855 ± 0.007	0.829 ± 0.008	0.835 ± 0.010
Computers	**0.691 ± 0.009**	0.690 ± 0.008	0.683 ± 0.011	0.587 ± 0.017	0.590 ± 0.007	0.510 ± 0.033	0.671 ± 0.006	0.588 ± 0.011	0.660 ± 0.010
Education	**0.624 ± 0.015**	0.623 ± 0.006	0.602 ± 0.009	0.479 ± 0.004	0.485 ± 0.002	0.457 ± 0.038	0.607 ± 0.009	0.481 ± 0.012	0.612 ± 0.011
Entertainment	**0.678 ± 0.018**	0.676 ± 0.012	0.659 ± 0.015	0.489 ± 0.005	0.535 ± 0.007	0.538 ± 0.020	0.646 ± 0.008	0.503 ± 0.014	0.667 ± 0.009
Health	0.750 ± 0.005	0.750 ± 0.010	0.746 ± 0.008	0.610 ± 0.007	0.737 ± 0.014	0.716 ± 0.013	**0.772 ± 0.009**	0.644 ± 0.011	0.698 ± 0.020
Recreation	0.598 ± 0.007	**0.602 ± 0.016**	0.589 ± 0.016	0.383 ± 0.007	0.422 ± 0.017	0.450 ± 0.022	0.563 ± 0.010	0.394 ± 0.007	0.591 ± 0.012
Reference	**0.682 ± 0.010**	0.679 ± 0.008	0.532 ± 0.006	0.557 ± 0.013	0.561 ± 0.006	0.536 ± 0.031	0.669 ± 0.004	0.558 ± 0.008	0.666 ± 0.018
Science	**0.561 ± 0.018**	0.560 ± 0.018	0.547 ± 0.009	0.388 ± 0.007	0.425 ± 0.013	0.422 ± 0.028	0.528 ± 0.012	0.406 ± 0.003	0.553 ± 0.011
Social	0.715 ± 0.012	**0.717 ± 0.009**	0.702 ± 0.009	0.616 ± 0.014	0.640 ± 0.005	0.585 ± 0.028	0.709 ± 0.008	0.625 ± 0.014	0.697 ± 0.015
Dataset	*Macro-averaging AUC* ↑								
	MGOFA	WRAP	HOMI	SLOFS	MDFS	TIFS	RLFSCL	GLFS	MC-GM
Art	**0.958 ± 0.010**	**0.958 ± 0.001**	0.957 ± 0.001	0.933 ± 0.002	0.936 ± 0.001	0.604 ± 0.022	0.956 ± 0.001	0.932 ± 0.002	0.957 ± 0.002
Business	**0.959 ± 0.001**	**0.959 ± 0.001**	**0.959 ± 0.001**	0.933 ± 0.007	0.925 ± 0.002	0.562 ± 0.013	0.957 ± 0.001	0.917 ± 0.008	0.956 ± 0.001
Computers	**0.969 ± 0.001**	**0.969 ± 0.001**	0.968 ± 0.001	0.945 ± 0.002	0.949 ± 0.002	0.594 ± 0.013	0.968 ± 0.000	0.940 ± 0.005	0.968 ± 0.002
Education	**0.971 ± 0.001**	0.970 ± 0.001	0.967 ± 0.003	0.950 ± 0.004	0.952 ± 0.002	0.594 ± 0.025	0.968 ± 0.003	0.946 ± 0.006	0.967 ± 0.002
Entertainment	**0.957 ± 0.001**	**0.957 ± 0.001**	0.956 ± 0.001	0.926 ± 0.002	0.942 ± 0.001	0.637 ± 0.014	0.956 ± 0.001	0.930 ± 0.010	0.956 ± 0.002
Health	0.965 ± 0.001	0.965 ± 0.002	0.965 ± 0.001	0.954 ± 0.007	0.965 ± 0.002	0.591 ± 0.006	**0.966 ± 0.001**	0.961 ± 0.000	0.965 ± 0.001
Recreation	**0.959 ± 0.001**	0.958 ± 0.002	0.958 ± 0.002	0.931 ± 0.002	0.940 ± 0.002	0.623 ± 0.013	0.956 ± 0.001	0.933 ± 0.001	0.957 ± 0.001
Reference	**0.977 ± 0.000**	0.975 ± 0.003	0.834 ± 0.017	0.953 ± 0.006	0.963 ± 0.002	0.614 ± 0.013	0.975 ± 0.002	0.954 ± 0.001	0.975 ± 0.001
Science	**0.976 ± 0.001**	**0.976 ± 0.001**	0.975 ± 0.001	0.959 ± 0.002	0.966 ± 0.001	0.615 ± 0.016	**0.976 ± 0.001**	0.958 ± 0.002	0.974 ± 0.002
Social	**0.978 ± 0.001**	0.977 ± 0.003	0.977 ± 0.001	0.959 ± 0.007	0.968 ± 0.002	0.627 ± 0.011	0.976 ± 0.003	0.956 ± 0.003	**0.978 ± 0.000**

4.5. Performance variations during convergence

Fig. 6 shows the performance variations on datasets *Art*, *Business*, and *Computers* with the continuously updating of $\mathbf{C}_r$, $\mathbf{V}_r$ and $\mathbf{b}_r$ for solving Eq. (13). Since the objective function converges with at most 400 iterations for the three datasets, we record the variations of Hamming Loss, Ranking Loss, One Error, Average Precision and Macro-averaging Precision every 10 iterations. As shown in Fig. 6, the performance variation becomes negligible for *Art* and *Business* when the iteration count reaches 100, while almost remains unchanged for *Computers* after 200 iterations. These observations imply that MGOFA converges simultaneously for both the objective function and classification performance.

4.6. Sensitivity analysis

To explore the performance fluctuations w.r.t. parameter settings, we conduct sensitivity analysis for MGOFA on datasets *Art*, *Business*,

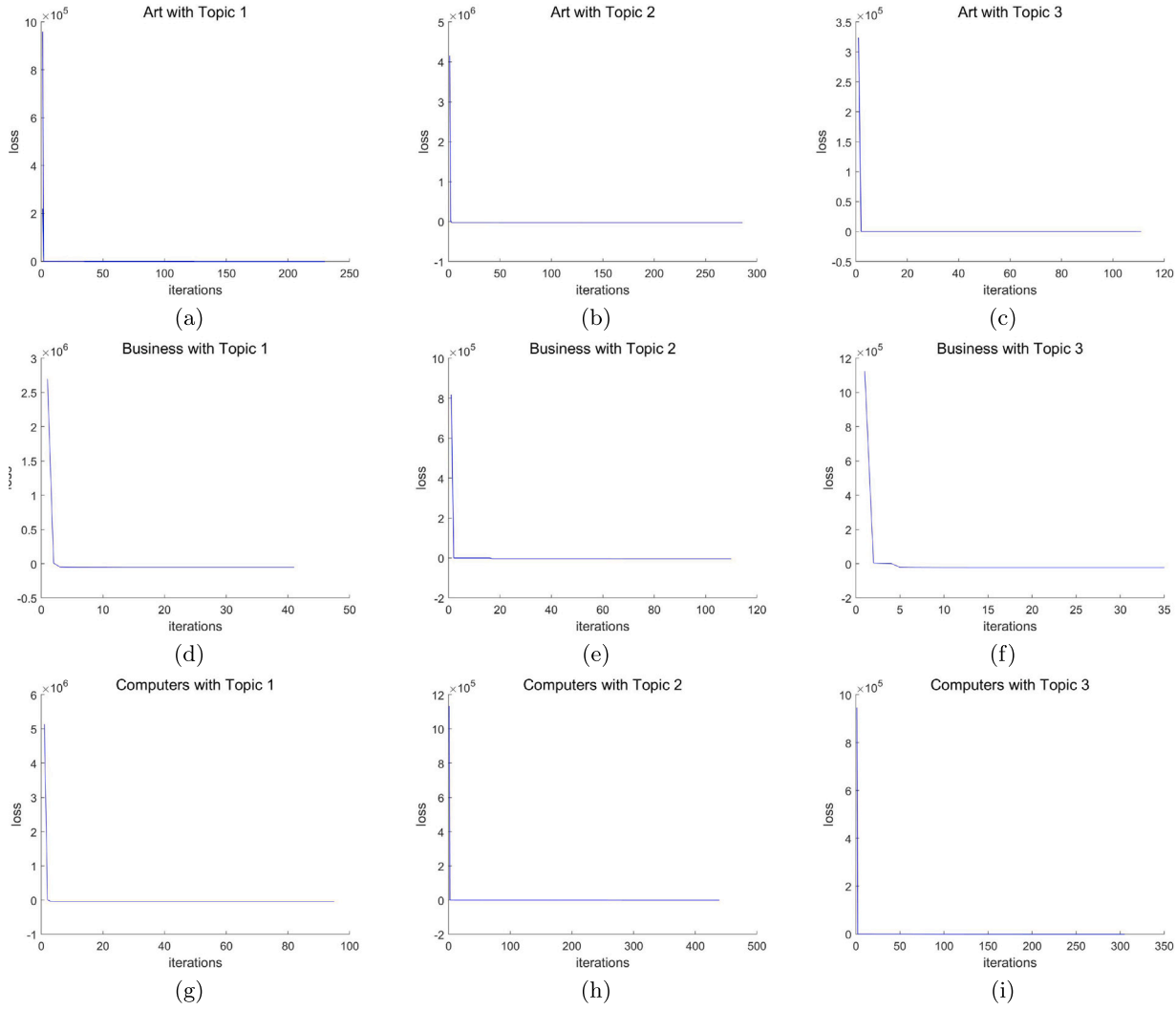


Fig. 5. Convergence curves on dataset, where *Art* for (a) Topic 1, (b) Topic 2, and (c) Topic 3, *Business* for (d) Topic 1, (e) Topic 2, and (f) Topic 3, and *Computers* for (g) Topic 1, (h) Topic 2, (i) Topic 3.

Table 5
Summary of the Friedman statistics F_F (N = 9, T = 10) and the critical value at significance level $\alpha = 0.05$ on all evaluation metrics.

Evaluation metric	F_F	Critical value
Hamming loss	88.1223	
Ranking loss	6.8404	
One error	53.1762	2.0698
Average precision	44.9191	
Macro − averaging AUC	44.5980	

and *Computers* over the parameters β, μ, k, t, λ_1, and λ_2.

We devise four groups of comparisons to examine the performance fluctuations from modules of granular-based feature augmentation, granular-based latent topic mining, and fine-grained label correlation learning. Details are as follows:

1. β: Parameter β is searched from 0.1 to 0.9 at a step of 0.1, while the remaining parameters $(k, \mu, t, \lambda_1, \lambda_2, \lambda_3)$ are fixed as $(10, 0.4, 3, 5, 5, 1)$.
2. (k, μ): Parameters μ and k are searched from 0.1 to 0.9 at a step of 0.1 and from 3 to 12 at a step of 1, respectively, while the remaining parameters $(\beta, t, \lambda_1, \lambda_2, \lambda_3)$ are fixed as $(0.9, 3, 5, 5, 1)$.
3. t: Parameter t is searched from 1 to 128 at a step of 2^n, where $n = 0, 1, \ldots, 6$, while the remaining parameters $(\beta, k, \mu, \lambda_1, \lambda_2, \lambda_3)$ are fixed as $(0.9, 10, 0.4, 5, 5, 1)$.
4. (λ_1, λ_2): Parameters λ_1 and λ_2 are searched from 1 to 10 at a step of 1, while the remaining parameters $(\beta, k, \mu, t, \lambda_3)$ are fixed as $(0.9, 10, 0.4, 3, 1)$.

4.6.1. Varying the dimensionality of embedding features

Fig. 7 shows the results of group 1 (i.e., (β)). It clearly reveals that the overall values of loss-based measures (i.e., *Hamming Loss*, *Ranking Loss*, and *One Error*) become smaller if β selects a larger value. The overall values of accuracy-based measures (i.e., *Average Precision*, *Macro-averaging AUC*) become larger if β selects a larger value. These observations suggest that a larger value of β can preserve more discriminative information.

4.6.2. Varying the trade-off weight μ and neighborhood size k

Fig. 8 shows the results of group 2 (i.e., (k, μ)). It shows the optimal performance is both dataset-dependent and metric-dependent. Typically, the results of *Art* become better if μ is medium while k is smaller; the results of *Business* become better if μ is larger while k is smaller; the results of *Computers* become better if both μ and k are smaller. These findings suggest that it is valuable to scrutinize the trade-off between

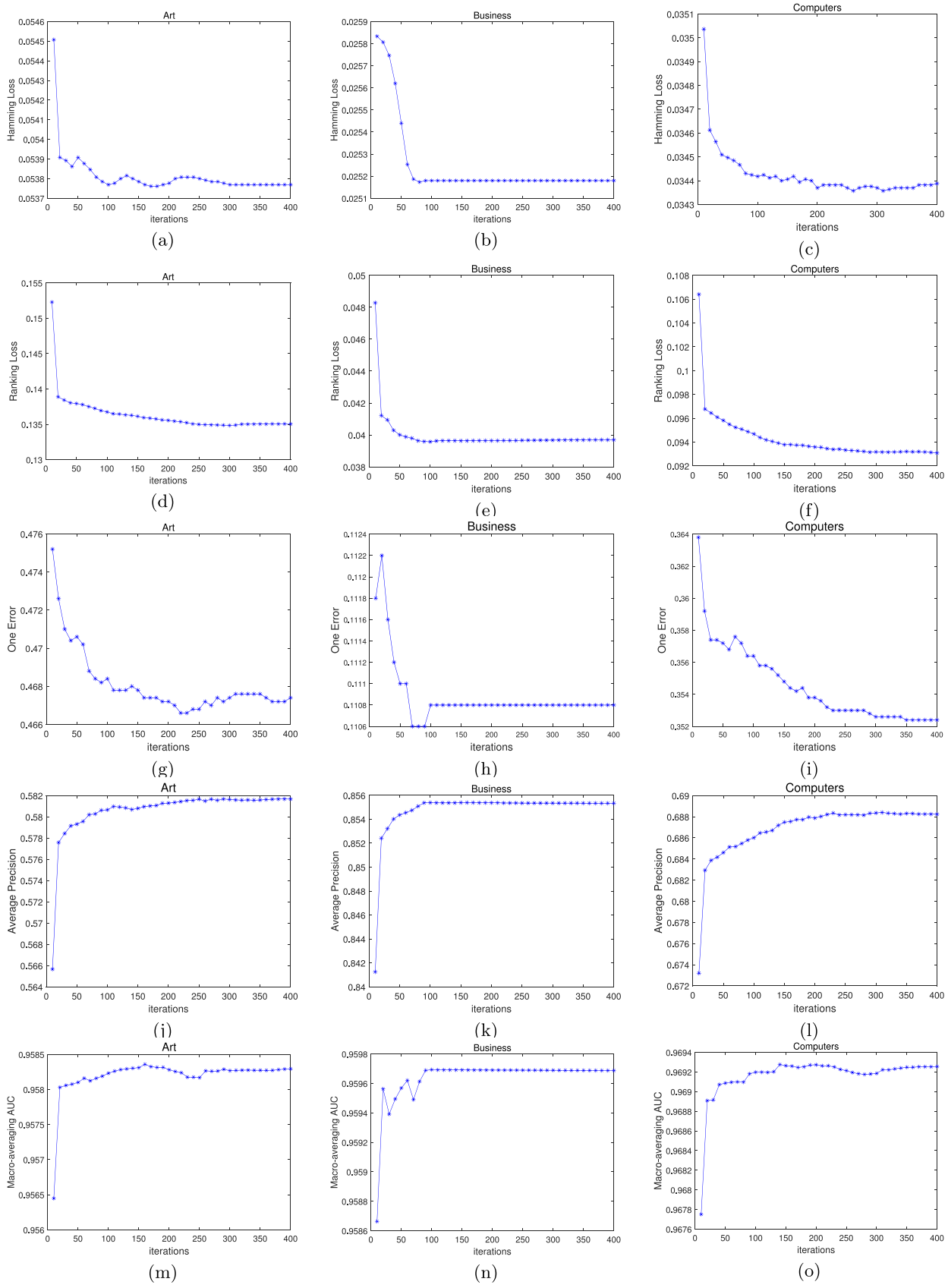


Fig. 6. Performance variations with the increment of iteration count for *Art* on (a), (d), (g), (j), (m), for *Business* on (b), (e), (h), (k), (n) and for *Computers* on (c), (f), (i), (l), (o).

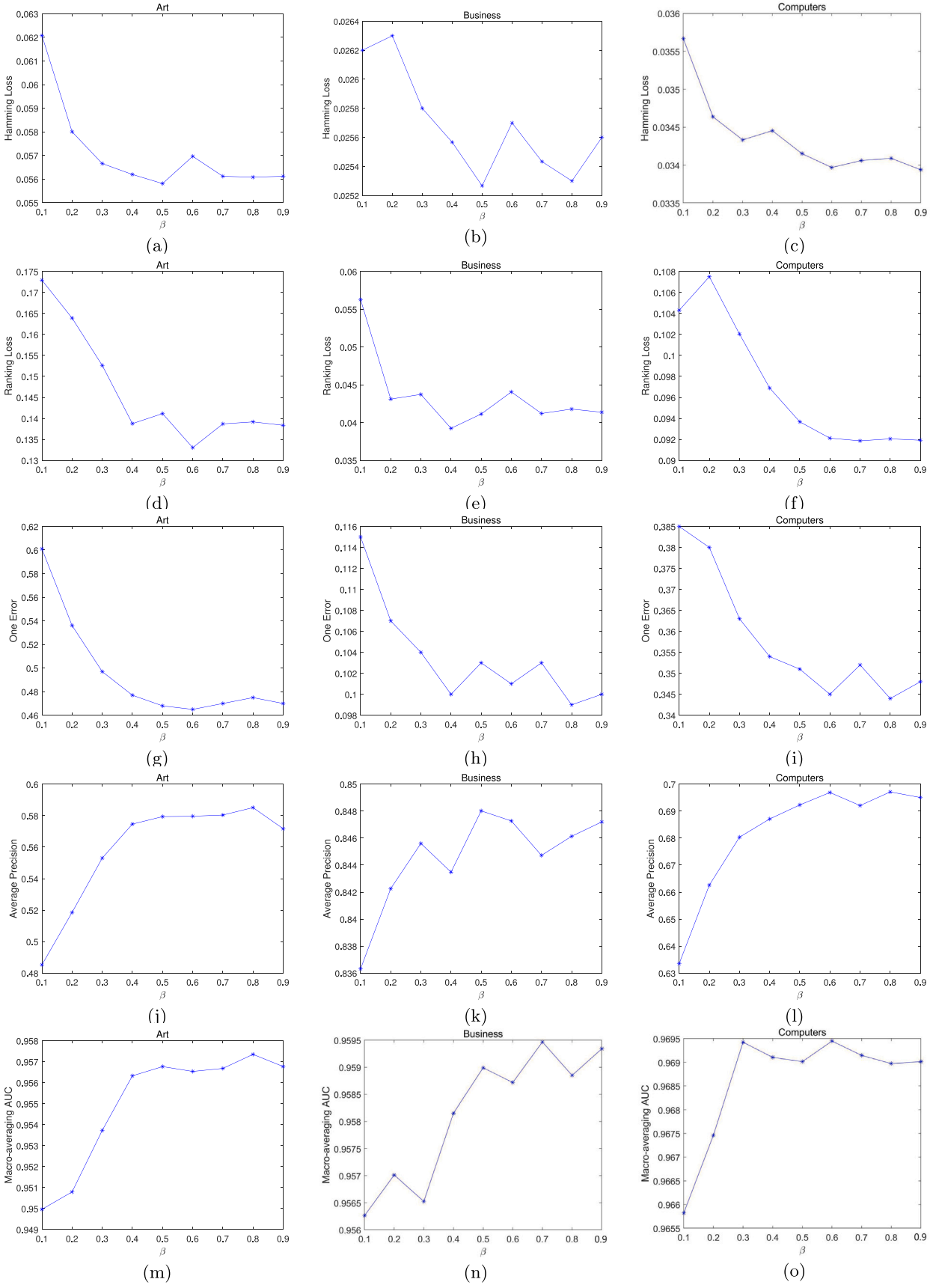


Fig. 7. Sensitivity analysis with varying parameter β for *Art* on (a), (d), (g), (j), (m), for *Business* on (b), (e), (h), (k), (n) and for *Computers* on (c), (f), (i), (l), (o).

Table 6
Comparisons of MGOFA (control algorithm) against the remaining algorithms with the Holm test at the significance level $\alpha = 0.05$. Algorithms that are statistically inferior to MGOFA are shown in bold size.

Hamming loss				
j	Algorithm	z_j	p_j	$\alpha/(k-j+1)$
2	**TIFS**	−5.837951	5.2847e−9	0.00625
3	**MC-GM**	−5.266403	1.3912e−7	0.00714
4	**SLOFS**	−4.123308	3.7347e−5	0.00833
5	**GLFS**	−3.674235	2.3856e−4	0.01000
6	**MDFS**	−3.021037	0.002519	0.01250
7	HOMI	−2.245366	0.024745	0.01667
8	RLFSCL	−1.102270	0.270344	0.02500
9	WRAP	−0.081650	0.934925	0.05000
Ranking loss				
j	Algorithm	z_j	p_j	$\alpha/(k-j+1)$
2	**TIFS**	−3.470110	5.2025e−4	0.00625
3	**SLOFS**	−3.388461	7.0286e−4	0.00714
4	**MC-GM**	−3.143512	0.0017	0.00833
5	GLFS	−2.204541	0.0275	0.01000
6	MDFS	−2.041241	0.0412	0.01250
7	RLFSCL	−1.592168	0.1130	0.01667
8	WRAP	−0.122474	0.9025	0.02500
9	HOMI	0.163299	1.0000	0.05000
One error				
j	Algorithm	z_j	p_j	$\alpha/(k-j+1)$
2	**SLOFS**	−5.307228	1.1130e−9	0.00625
3	**TIFS**	−5.103104	3.3413e−7	0.00714
4	**GLFS**	−4.73568	2.1832e−6	0.00833
5	**MDFS**	−4.082483	4.4557e−5	0.01000
6	RLFSCL	−2.245366	0.024745	0.01250
7	HOMI	−2.204541	0.027486	0.01667
8	MC-GM	−1.551344	0.120819	0.02500
9	WRAP	−0.122474	0.902523	0.05000
Average precision				
j	Algorithm	z_j	p_j	$\alpha/(k-j+1)$
2	**SLOFS**	−5.715476	1.0940e−8	0.00625
3	**TIFS**	−4.980629	6.3378e−7	0.00714
4	**GLFS**	−4.81733	1.4549e−6	0.00833
5	**MDFS**	−3.919184	8.8849e−5	0.01000
6	HOMI	−2.44949	0.014306	0.01250
7	MC-GM	−2.28619	0.022243	0.01667
8	RLFSCL	−1.918767	0.055014	0.02500
9	WRAP	−0.367423	0.713303	0.05000
Macro-averaging AUC				
j	Algorithm	z_j	p_j	$\alpha/(k-j+1)$
2	**TIFS**	−5.960425	2.5158e−9	0.00625
3	**GLFS**	−4.73568	4.2948e−8	0.00714
4	**SLOFS**	−4.490731	7.0979e−6	0.00833
5	**MDFS**	−3.347636	8.1504e−4	0.01000
6	HOMI	−1.877942	0.060389	0.01250
7	MC-GM	−1.755468	0.079179	0.01667
8	RLFSCL	−1.551344	0.120819	0.02500
9	WRAP	−0.530723	0.530723	0.05000

the Pearson correlation coefficient and cosine similarity for local label correlation. Furthermore, having a large-scale neighborhood may not necessarily guarantee a better outcome for cases with underlying topics.

4.6.3. Varying the number of topics t

Fig. 9 shows the results of group 3 (i.e., (t)). They obviously illustrate that the latent topics in *Art*, *Business*, and *Computers* are limited in scale. Concretely, the classification performance is rapidly degenerated if too many topics are assumed. It implies the poor generalization ability of label correlation in a few instances. Consequently, a reasonable number of topics is crucial in boosting classification performance.

4.6.4. Varying the penalties λ_1 and λ_2

Fig. 10 shows the results of group 4 (i.e., (λ_1, λ_2)). The fluctuations incurred by variations from λ_1 and λ_2 are insignificant. It implies that MGOFA is insensitive to the variations of λ_1 and λ_2.

4.7. Ablation study

To explore the functionality of label correlation and feature augmentation in MGOFA, four reduced versions of MGOFA are studied:

1. MGOFA-LC: A reduced version of MGOFA without leveraging label correlation. This means that the regularization term $\frac{\lambda_1}{2} tr\left(\mathbf{U}_r \mathbf{C}_r \mathbf{U}_r^\top\right)$ in Eq. (13) is removed.
2. MGOFA-FA: A reduced version of MGOFA without exploring augmented features. The fine-grained label correlation learning in this version learns the topic-based label correlation on the m-dimensional feature space.
3. MGOFA-FS: A reduced version of MGOFA without imposing sparsity constraints on the weights of features. This means that the regularization term $\lambda_3 \|\mathbf{U}\|_1$ in Eq. (13) is removed.
4. MGOFA-FE: A reduced version of MGOFA without learning feature embedding. This means that both the regularization term $\frac{\lambda_2}{2}\|\mathbf{V}_r\|_2^2$ and the construction of $\mathbf{V}_r$ in Eq. (13) are removed.

Table 7 shows the results of the ablation study. The smaller the ranking is, the better the classification performance becomes. It clearly shows that all these components contribute to the effectiveness of MGOFA, as MGOFA dominates the MGOFA-LC (with the win/tie/lose as 44/6/0), MGOFA-FA (with the win/tie/lose as 48/2/0), MGOFA-FS (with the win/tie/lose as 28/22/0), and MGOFA-FE (with the win/tie/lose as 50/0/0). For the degenerated versions, the larger ranking implies the more significant contributions to the MGOFA. Considering the ranking among the MGOFA and all degenerated versions, the importance of the four modules is "Feature Embedding > Feature Augmentation > Label Correlation > Feature Sparsity". Furthermore, the functionality of augmented features is more important than the label correlation, demonstrating that the likelihood of bias incurred by augmented features is negligible.

4.8. Computational efficiency evaluation

Table 8 provides a comparison of the computational time required for different approaches. Parameter settings for MGOFA are $(\beta, k, \mu, t, \lambda_1, \lambda_2, \lambda_3) = (0.9, 10, 0.4, 3, 5, 5, 1)$, while the comparing algorithms take the settings as declared in Section 4.2. As shown in Table 8, SLOFS, which employs a filtering-based feature selection strategy, emerges as the most efficient solution. In contrast, the proposed MGOFA outperforms approximately half of the state-of-the-art methods. Given that the primary objective is effectiveness, the computational efficiency ranking of MGOFA is acceptable.

5. Conclusions

We present a multi-granularity correlation-based feature augmentation (MGOFA) method to deduce fine-grained local label correlations for multi-label classification. With the semantics of relative positive tendency towards label space and a limited number of such features, we bridge the gap between feature representation and classification by concatenating the low-level original features with the high-level augmented features for each topic independently. By learning the mapping from the multi-faceted information to the ground-truth labels in a wrapped manner, we demonstrate that it not only achieves the statistical superiority performance over the state-of-the-art approaches, but also converges rapidly and fluctuates stably.

The MGOFA demonstrates it is feasible to employ multi-granularity on multi-label classification in dealing with the uncertainty of local

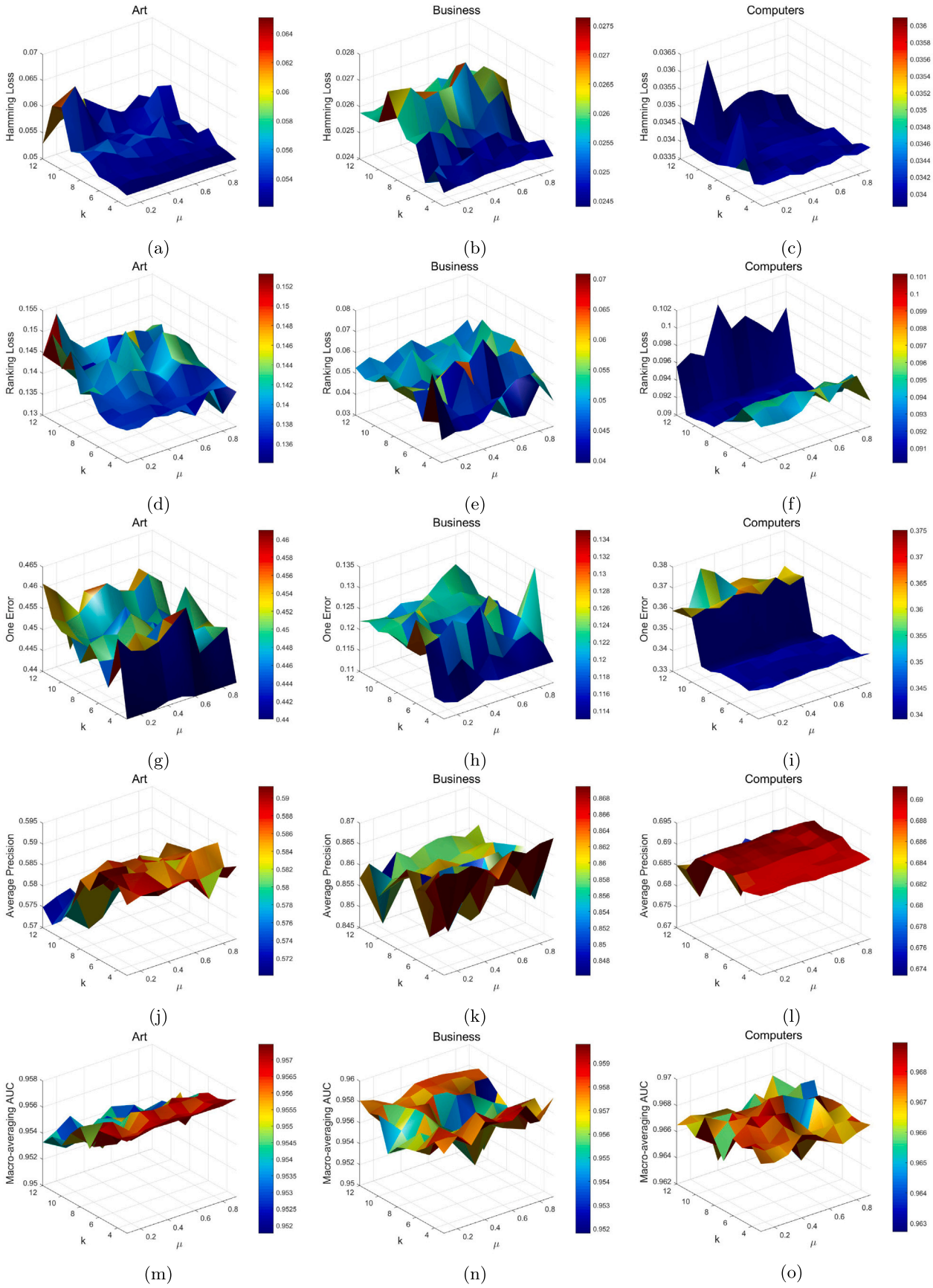


Fig. 8. Sensitivity analysis with varying parameters k and μ for *Art* on (a), (d), (g), (j), (m), for *Business* on (b), (e), (h), (k), (n) and for *Computers* on (c), (f), (i), (l), (o).

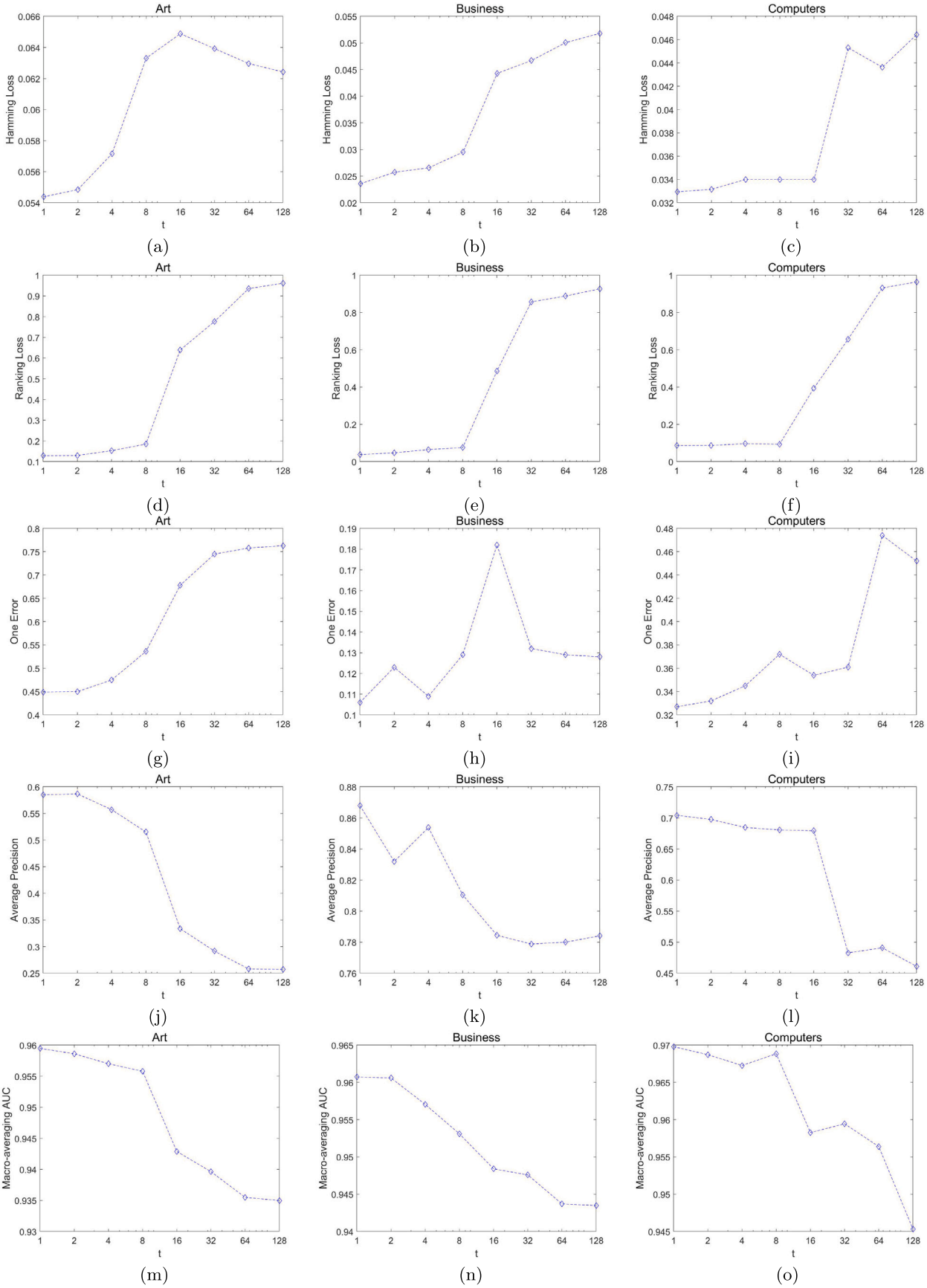


Fig. 9. Sensitivity analysis with varying parameter t for *Art* on (a), (d), (g), (j), (m), for *Business* on (b), (e), (h), (k), (n) and for *Computers* on (c), (f), (i), (l), (o).

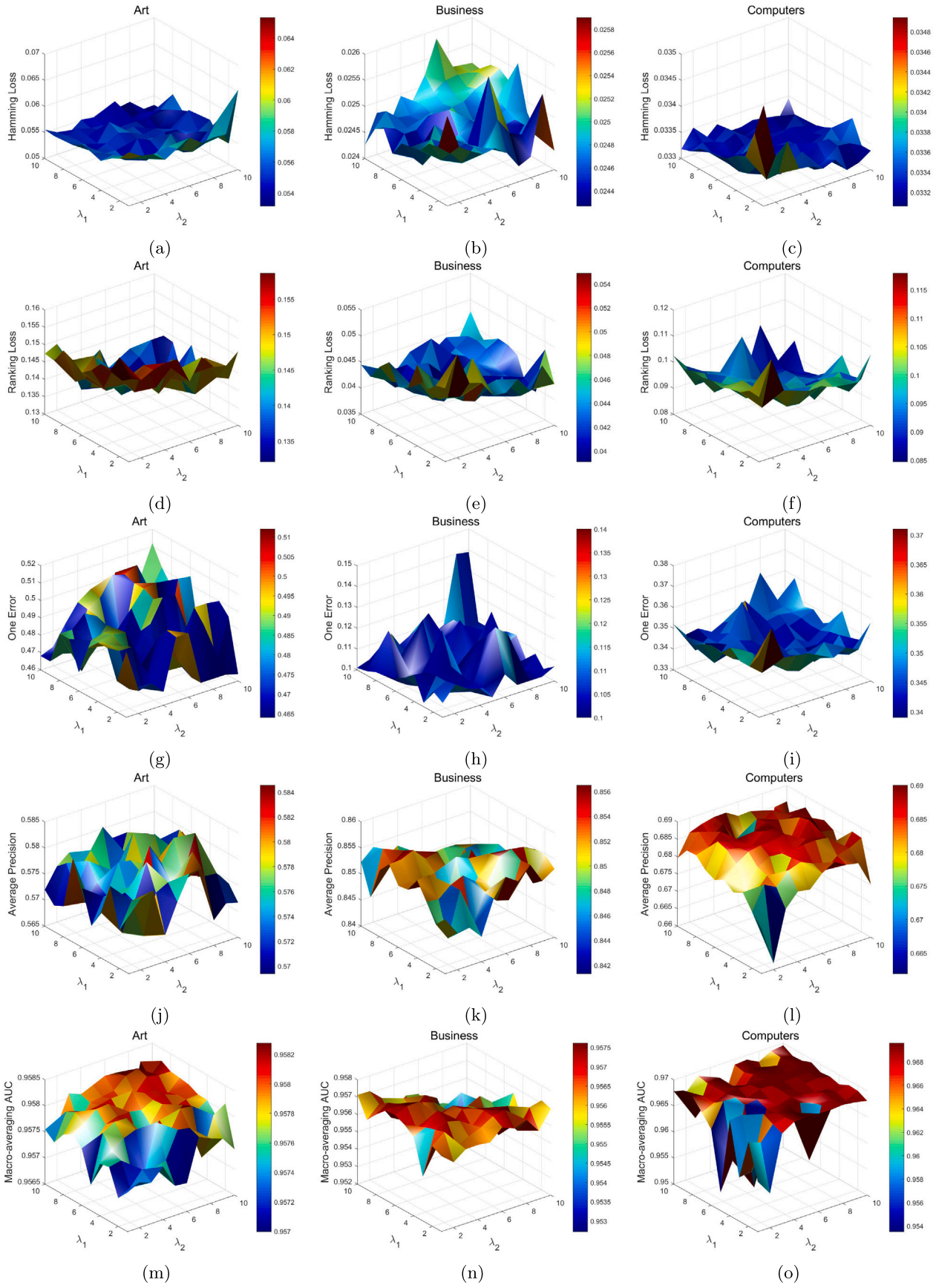


Fig. 10. Sensitivity analysis with varying parameters λ_1 and λ_2 for *Art* on (a), (d), (g), (j), (m), for *Business* on (b), (e), (h), (k), (n) and for *Computers* on (c), (f), (i), (l), (o).

Table 7
Functionality analysis on MGOFA (↓: the smaller the better, ↑: the larger the better).

Dataset	*Hamming loss* ↓				
	MGOFA	MGOFA-LC	MGOFA-FA	MGOFA-FS	MGOFA-FE
Art	**0.054 ± 0.001**(2)	**0.054 ± 0.001**(2)	0.057 ± 0.000(5)	**0.054 ± 0.001**(2)	0.056 ± 0.045(4)
Business	**0.025 ± 0.000**(1.5)	0.026 ± 0.001(3.5)	0.026 ± 0.001(3.5)	**0.025 ± 0.001**(1.5)	0.028 ± 0.006(5)
Computers	**0.034 ± 0.001**(1.5)	0.036 ± 0.002(4.5)	0.035 ± 0.001(3)	**0.034 ± 0.001**(1.5)	0.036 ± 0.067(4.5)
Education	**0.037 ± 0.001**(1.5)	0.038 ± 0.000(3)	0.040 ± 0.002(4)	**0.037 ± 0.001**(1.5)	0.044 ± 0.126(5)
Entertainment	**0.052 ± 0.002**(1.5)	0.053 ± 0.001(3.5)	0.054 ± 0.003(5)	**0.052 ± 0.001**(1.5)	0.053 ± 0.003(3.5)
Health	**0.035 ± 0.001**(2.5)	**0.035 ± 0.001**(2.5)	**0.035 ± 0.001**(2.5)	**0.035 ± 0.001**(2.5)	0.044 ± 0.020(5)
Recreation	**0.054 ± 0.001**(1.5)	0.055 ± 0.001(3.5)	0.058 ± 0.002(5)	**0.054 ± 0.001**(1.5)	0.055 ± 0.004(3.5)
Reference	**0.026 ± 0.001**(2)	**0.026 ± 0.001**(2)	0.028 ± 0.001(4)	**0.026 ± 0.001**(2)	0.031 ± 0.002(5)
Science	**0.032 ± 0.001**(1.5)	0.034 ± 0.002(4)	0.035 ± 0.003(5)	**0.032 ± 0.001**(1.5)	0.033 ± 0.078(3)
Social	**0.021 ± 0.001**(1)	0.026 ± 0.002(5)	0.023 ± 0.002(4)	0.022 ± 0.001(2.5)	0.022 ± 0.077(2.5)
Dataset	*Ranking loss* ↓				
	MGOFA	MGOFA-LC	MGOFA-FA	MGOFA-FS	MGOFA-FE
Art	**0.135 ± 0.005**(1)	0.149 ± 0.005(3)	0.162 ± 0.005(5)	0.136 ± 0.005(2)	0.151 ± 0.020(4)
Business	**0.040 ± 0.003**(1)	0.048 ± 0.006(4.5)	0.045 ± 0.004(3)	0.041 ± 0.005(2)	0.048 ± 0.017(4.5)
Computers	**0.092 ± 0.009**(1.5)	0.109 ± 0.003(4)	0.112 ± 0.010(5)	**0.092 ± 0.004**(1.5)	0.094 ± 0.015(3)
Education	**0.106 ± 0.008**(1.5)	0.121 ± 0.008(3)	0.132 ± 0.017(4)	**0.106 ± 0.006**(1.5)	0.143 ± 0.013(5)
Entertainment	**0.113 ± 0.007**(1.5)	0.126 ± 0.007(4)	0.127 ± 0.015(5)	**0.113 ± 0.007**(1.5)	0.125 ± 0.007(3)
Health	**0.062 ± 0.004**(1)	0.071 ± 0.003(4)	0.065 ± 0.003(3)	0.064 ± 0.003(2)	0.111 ± 0.005(5)
Recreation	**0.149 ± 0.004**(1)	0.162 ± 0.008(4)	0.171 ± 0.005(5)	0.151 ± 0.007(2)	0.154 ± 0.015(3)
Reference	**0.093 ± 0.006**(1.5)	0.102 ± 0.006(3)	0.117 ± 0.010(4)	**0.093 ± 0.008**(1.5)	0.122 ± 0.003(5)
Science	**0.134 ± 0.010**(1)	0.149 ± 0.004(3)	0.152 ± 0.010(4)	0.136 ± 0.007(2)	0.156 ± 0.025(5)
Social	**0.082 ± 0.004**(1)	0.103 ± 0.006(4)	0.097 ± 0.014(3)	0.083 ± 0.003(2)	0.114 ± 0.015(5)
Dataset	*One error* ↓				
	MGOFA	MGOFA-LC	MGOFA-FA	MGOFA-FS	MGOFA-FE
Art	**0.460 ± 0.010**(1)	0.468 ± 0.016(3)	0.493 ± 0.021(5)	0.464 ± 0.013(2)	0.485 ± 0.011(4)
Business	**0.111 ± 0.004**(1)	0.112 ± 0.010(3)	0.112 ± 0.009(3)	0.112 ± 0.008(3)	0.116 ± 0.021(5)
Computers	**0.353 ± 0.018**(1)	0.361 ± 0.007(3.5)	0.367 ± 0.005(5)	0.354 ± 0.009(2)	0.361 ± 0.101(3.5)
Education	**0.468 ± 0.018**(1)	0.476 ± 0.014(3)	0.497 ± 0.017(4)	0.474 ± 0.010(2)	0.513 ± 0.031(5)
Entertainment	**0.399 ± 0.017**(1)	0.415 ± 0.006(3)	0.422 ± 0.021(5)	0.406 ± 0.012(2)	0.421 ± 0.006(4)
Health	**0.273 ± 0.010**(1)	0.274 ± 0.012(2.5)	0.275 ± 0.014(4)	0.274 ± 0.011(2.5)	0.342 ± 0.011(5)
Recreation	**0.461 ± 0.006**(1)	0.476 ± 0.016(3)	0.496 ± 0.014(5)	0.472 ± 0.016(2)	0.486 ± 0.006(4)
Reference	**0.373 ± 0.010**(1)	0.377 ± 0.012(3)	0.404 ± 0.006(4)	0.376 ± 0.021(2)	0.427 ± 0.004(5)
Science	**0.499 ± 0.019**(1)	0.514 ± 0.013(3)	0.518 ± 0.017(4)	0.511 ± 0.014(2)	0.535 ± 0.024(5)
Social	**0.294 ± 0.015**(1)	0.320 ± 0.010(4)	0.305 ± 0.027(3)	0.303 ± 0.012(2)	0.320 ± 0.076(5)
Dataset	*Average precision* ↑				
	MGOFA	MGOFA-LC	MGOFA-FA	MGOFA-FS	MGOFA-FE
Art	**0.583 ± 0.008**(1)	0.573 ± 0.009(3)	0.558 ± 0.014(5)	0.580 ± 0.008(2)	0.564 ± 0.019(4)
Business	**0.856 ± 0.006**(1)	0.845 ± 0.014(4)	0.847 ± 0.007(3)	0.854 ± 0.013(2)	0.838 ± 0.018(5)
Computers	**0.691 ± 0.009**(1)	0.676 ± 0.003(4)	0.674 ± 0.011(5)	0.689 ± 0.004(3)	0.690 ± 0.076(2)
Education	**0.624 ± 0.015**(1)	0.611 ± 0.010(3)	0.598 ± 0.018(4)	0.620 ± 0.006(2)	0.580 ± 0.048(5)
Entertainment	**0.678 ± 0.018**(1)	0.663 ± 0.007(3)	0.661 ± 0.024(5)	0.674 ± 0.009(2)	0.662 ± 0.005(4)
Health	**0.750 ± 0.005**(1)	0.749 ± 0.006(2)	0.745 ± 0.004(4)	0.748 ± 0.007(3)	0.681 ± 0.011(5)
Recreation	**0.598 ± 0.007**(1)	0.584 ± 0.013(3)	0.568 ± 0.014(5)	0.592 ± 0.013(2)	0.579 ± 0.007(4)
Reference	**0.682 ± 0.010**(1)	0.671 ± 0.012(3)	0.656 ± 0.008(4)	0.681 ± 0.017(2)	0.634 ± 0.004(5)
Science	**0.561 ± 0.018**(1)	0.546 ± 0.007(3)	0.541 ± 0.015(4)	0.556 ± 0.007(2)	0.532 ± 0.034(5)
Social	**0.715 ± 0.012**(1)	0.688 ± 0.007(5)	0.702 ± 0.023(3)	0.710 ± 0.007(2)	0.690 ± 0.080(4)
Dataset	*Macro-averaging AUC* ↑				
	MGOFA	MGOFA-LC	MGOFA-FA	MGOFA-FS	MGOFA-FE
Art	**0.958 ± 0.010**(2)	**0.958 ± 0.001**(2)	0.956 ± 0.001(5)	**0.958 ± 0.001**(2)	0.957 ± 0.003(4)
Business	**0.959 ± 0.001**(2.5)	**0.959 ± 0.001**(2.5)	**0.959 ± 0.002**(2.5)	**0.959 ± 0.000**(2.5)	0.957 ± 0.001(5)
Computers	**0.969 ± 0.001**(1.5)	0.968 ± 0.001(3.5)	0.967 ± 0.001(5)	**0.969 ± 0.002**(1.5)	0.968 ± 0.003(3.5)
Education	**0.971 ± 0.001**(1)	0.969 ± 0.002(3.5)	0.969 ± 0.001(3.5)	0.970 ± 0.001(2)	0.968 ± 0.002(5)
Entertainment	**0.957 ± 0.001**(2)	**0.957 ± 0.001**(2)	0.955 ± 0.002(4)	**0.957 ± 0.001**(2)	0.944 ± 0.003(5)
Health	**0.965 ± 0.001**(1.5)	0.964 ± 0.001(3.5)	0.964 ± 0.001(3.5)	**0.965 ± 0.001**(1.5)	0.958 ± 0.002(5)
Recreation	**0.959 ± 0.001**(1.5)	0.958 ± 0.001(3.5)	0.956 ± 0.001(5)	**0.959 ± 0.001**(1.5)	0.958 ± 0.005(3.5)
Reference	**0.977 ± 0.000**(1.5)	0.975 ± 0.001(4)	0.975 ± 0.001(4)	**0.977 ± 0.001**(1.5)	0.975 ± 0.001(4)
Science	**0.976 ± 0.001**(1.5)	0.975 ± 0.002(4)	0.975 ± 0.001(4)	**0.976 ± 0.001**(1.5)	0.975 ± 0.001(4)
Social	**0.978 ± 0.001**(1.5)	0.977 ± 0.001(3.5)	0.977 ± 0.001(3.5)	**0.978 ± 0.001**(1.5)	0.974 ± 0.002(5)

Table 8
Comparisons (mean ± std) on time consumption.

Dataset	*Time consumption (unit:s)* ↓								
	MGOFA	WRAP	HOMI	SLOFS	MDFS	TIFS	RLFSCL	GLFS	MC-GM
Art	15.48 ± 2.143	5.116 ± 2.670	114.5 ± 0.733	**2.366 ± 0.251**	39.07 ± 3.434	81.32 ± 8.941	16.24 ± 0.272	12.56 ± 2.230	98.42 ± 0.175
Business	7.306 ± 1.920	**1.161 ± 0.082**	114.0 ± 0.570	2.139 ± 0.044	37.52 ± 0.967	74.60 ± 2.269	16.98 ± 0.222	10.00 ± 2.998	99.07 ± 0.189
Computers	48.78 ± 3.034	6.774 ± 3.574	137.1 ± 5.689	**2.422 ± 0.015**	39.67 ± 0.840	82.77 ± 3.026	30.67 ± 0.527	10.93 ± 1.981	101.7 ± 1.462
Education	13.36 ± 1.605	6.177 ± 0.688	123.0 ± 0.173	**2.299 ± 0.042**	46.18 ± 6.256	75.43 ± 0.458	23.48 ± 0.343	11.92 ± 2.384	99.19 ± 1.687
Entertainment	20.56 ± 1.230	7.027 ± 4.194	139.3 ± 3.133	**2.359 ± 0.030**	37.24 ± 0.685	78.57 ± 0.334	23.13 ± 0.347	16.19 ± 6.788	94.85 ± 0.481
Health	19.28 ± 3.404	2.668 ± 0.258	137.5 ± 1.125	**2.274 ± 0.021**	48.23 ± 0.337	84.63 ± 4.539	25.66 ± 0.513	13.47 ± 3.418	96.57 ± 1.753
Recreation	25.70 ± 4.144	8.623 ± 3.649	129.2 ± 3.035	**2.426 ± 0.080**	37.24 ± 0.532	85.98 ± 1.785	21.31 ± 0.301	16.18 ± 5.489	93.66 ± 0.293
Reference	38.05 ± 5.801	18.28 ± 4.977	206.0 ± 0.588	**2.623 ± 0.036**	41.74 ± 2.701	101.1 ± 8.209	37.68 ± 1.550	19.74 ± 15.76	98.53 ± 0.243
Science	51.49 ± 11.22	42.27 ± 15.78	141.7 ± 2.339	**2.566 ± 0.019**	30.38 ± 0.299	91.73 ± 1.952	41.91 ± 0.529	13.62 ± 2.161	96.40 ± 0.337
Social	85.11 ± 17.96	48.21 ± 21.54	177.2 ± 0.696	**3.065 ± 0.032**	44.13 ± 2.335	109.7 ± 9.657	64.73 ± 0.621	15.21 ± 1.421	97.94 ± 0.462
Average ranking	4.7(5)	2.4(2)	9(9)	1.1(1)	5.3(6)	7.2(7)	4.6(4)	2.9(3)	7.8(8)

label correlations. However, a shortcoming is the degenerated functionality of augmented features for insufficient supervision information cases. Therefore, we intend to extend the model to weakly supervised cases.

CRediT authorship contribution statement

Tianna Zhao: Writing – original draft, Visualization, Software, Investigation, Formal analysis, Conceptualization. **Yuanjian Zhang:** Writing – review & editing, Validation, Methodology, Conceptualization. **Duoqian Miao:** Supervision, Funding acquisition. **Witold Pedrycz:** 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.

Acknowledgments

Authors would like to thank the anonymous reviewers for their constructive comments and valuable suggestions. This work is supported by the National Key Research and Development Program of China (Grant No. 2022YFB3104702), and the National Natural Science Foundation of China (Grant No. 62376198).

Data availability

Data will be made available on request.

References

[1] M.L. Zhang, Z.H. Zhou, A review on multi-label learning algorithms, IEEE Trans. Knowl. Data Eng. 26 (8) (2014) 1819–1837.
[2] W.W. Liu, H.B. Wang, X.B. Shen, et al., The emerging trends of multi-label learning, IEEE Trans. Pattern Anal. Mach. Intell. 44 (11) (2022) 7955–7974.
[3] X. Kang, X.F. Shi, Y.N. Wu, et al., Active learning with complementary sampling for instructing class-biased multi-label text emotion classification, IEEE Trans. Affect. Comput. 14 (1) (2023) 523–536.
[4] F.G. Fan, Y.T. Su, L.Q. Nie, et al., Dual-domain aligned deep hierarchical matrix factorization method for micro–video multi-label classification, IEEE Trans. MultiMedia 26 (2024) 2598–2607.
[5] L. Ju, Z. Yu, L. Wang, et al., Hierarchical knowledge guided learning for real-world retinal disease recognition, IEEE Trans. Med. Imaging 43 (1) (2024) 335–350.
[6] G. Tsoumakas, I. Katakis, I. Vlahavas, Random *k*-labelsets for multilabel classification, IEEE Trans. Knowl. Data Eng. 23 (7) (2011) 1079–1089.
[7] M.L. Zhang, L. Wu, LIFT: Multi-label learning with label-specific features, IEEE Trans. Pattern Anal. Mach. Intell. 37 (1) (2015) 107–120.
[8] J. Huang, G.R. Li, Q.M. Huang, et al., Learning label-specific features and class-dependent labels for multi-label classification, IEEE Trans. Knowl. Data Eng. 28 (12) (2016) 3309–3323.
[9] C.Q. Zhang, Z.W. Yu, H.Z. Fu, et al., Hybrid noise-oriented multilabel learning, IEEE Trans. Cybern. 50 (6) (2020) 2837–2850.
[10] S.J. Huang, Z.H. Zhou, Multi-label learning by exploiting label correlations locally, in: Proceedings of 26th AAAI Conference on Artificial Intelligence, 2012, pp. 1–9.
[11] X.Y. Jia, Z.C. Li, X. Zheng, et al., Label distribution learning with label correlations on local samples, IEEE Trans. Knowl. Data Eng. 33 (4) (2021) 1619–1631.
[12] J.H. Ma, T.W.S. Chow, Latent topic-aware multioutput learning, IEEE Trans. Syst. Man Cybern.: Syst. 53 (12) (2023) 7547–7559.
[13] X.Y. Che, D.G. Chen, J.S. Mi, Learning instance-level label correlation distribution for multilabel classification with fuzzy rough sets, IEEE Trans. Fuzzy Syst. 31 (8) (2023) 2871–2884.
[14] Y. Zhu, J.T. Kwok, Z.H. Zhou, Multi-label learning with global and local label correlation, IEEE Trans. Knowl. Data Eng. 30 (6) (2018) 1081–1094.
[15] J. Zhang, Z.M. Luo, C.D. Li, et al., Manifold regularized discriminative feature selection for multi-label learning, Pattern Recognit. 95 (2019) 136–150.
[16] J.H. Ma, T.W.S. Chow, Topic-based instance and feature selection in multilabel classification, IEEE Trans. Neural Netw. Learn. Syst. 33 (1) (2022) 315–329.
[17] J.H. Ma, H.J. Zhang, T.W.S. Chow, Multilabel classification with label-specific features and classifiers: A coarse- and fine-tuned framework, IEEE Trans. Cybern. 51 (2) (2021) 1028–1042.
[18] J. Hobbs, Granularity, in: Proceedings of the 9th International Joint Conference on Artificial Intelligence, 1985, pp. 432–435.
[19] J.D. Qin, L. Martinez, W. Pedrycz, et al., An overview of granular computing in decision-making: Extensions, applications, and challenges, Inf. Fusion 98 (2023) 101833.
[20] Y.J. Zhang, T.N. Zhao, D.Q. Miao, et al., Granular multilabel batch active learning with pairwise label correlation, IEEE Trans. Syst. Man Cybern.: Syst. 52 (5) (2022) 3079–3091.
[21] X.Y. Che, D.G. Chen, J. Deng, et al., Exploiting local label correlation from sample perspective for multi-label classification via three-way decision theory, Appl. Soft Comput. 149 (2023) 110950.
[22] Z. Qin, H.M. Chen, Y. Mi, et al., Multi-label feature selection with adaptive graph learning and label information enhancement, Knowl.-Based Syst. 285 (2024) 111363.
[23] J.M. Wu, E.C.C. Tsang, W.H. Xu, et al., Correlation concept-cognitive learning model for multi-label classification, Knowl.-Based Syst. 290 (2024) 111566.
[24] T.N. Zhao, Y.J. Zhang, D.Q. Miao, Granular correlation-based label-specific feature augmentation for multi-label classification, Inform. Sci. 689 (2025) 121473.
[25] Y.J. Zhang, T.N. Zhao, D.Q. Miao, Y.Y. Yao, Three-way multi-label classification: a review, a framework, and new challenges, Appl. Soft Comput. 171 (2025) 112757.
[26] B.B. Jia, M.L. Zhang, Multi-dimensional classification via *k*NN feature augmentation, Pattern Recognit. 106 (2020) 107423.
[27] B.B. Jia, M.L. Zhang, Multi-dimensional classification via selective feature augmentation, Mach. Intell. Res. 19 (1) (2022) 38–51.
[28] Z.B. Yu, M.L. Zhang, Multi-label classification with label-specific feature generation: A wrapped approach, IEEE Trans. Pattern Anal. Mach. Intell. 44 (9) (2022) 5199–5210.
[29] J. Huang, G.R. Li, S.H. Wang, et al., Multi-label classification by exploiting local positive and negative pairwise label correlation, Neurocomput. 257 (2017) 164–174.
[30] W. Weng, Y.J. Lin, S.X. Wu, et al., Multi-label learning based on label-specific features and local pairwise label correlation, Neurocomput. 273 (2018) 385–394.
[31] W. Pedrycz, Evaluating quality of models via prediction information granules, IEEE Trans. Fuzzy Syst. 30 (12) (2022) 5551–5556.
[32] D.Y. Xia, G.Y. Wang, J. Yang, et al., Local knowledge distance for rough approximation measure in multi-granularity spaces, Inform. Sci. 605 (2022) 413–432.

[33] Y. Wang, Q.H. Hu, H. Chen, et al., Uncertainty instructed multi-granularity decision for large-scale hierarchical classification, Inform. Sci. 586 (2022) 644–661.
[34] X.Y. Zhang, H.Y. Gou, Statistical-mean double-quantitative k-nearest neighbor classification learning based on neighborhood distance measurement, Knowl.-Based Syst. 250 (2022) 109018.
[35] X. Shu, L. Zhang, Z.Z. Wang, et al., Fine-grained recognition: Multi-granularity labels and category similarity matrix, Knowl.-Based Syst. 273 (2023) 110599.
[36] B. Yu, H.J. Xie, M.J. Cai, et al., MG-GCN: Multi-granularity graph convolutional neural network for multi-label classification in multi-label information system, IEEE Trans. Emerg. Top. Comput. Intell. 8 (1) (2024) 288–299.
[37] Y. Yu, M. Wan, J. Qian, et al., Feature selection for multi-label learning based on variable-degree multi-granulation decision-theoretic rough sets, Internat. J. Approx. Reason. 169 (2024) 109181.
[38] L. Antwarg, C. Galed, N. Shimoni, et al., Shapley-based feature augmentation, Inf. Fusion 96 (2023) 92–102.
[39] H.Y. Zhang, M. Li, D.Q. Miao, et al., Construction of a feature enhancement network for small object detection, Pattern Recognit. 143 (2023) 109801.
[40] Z.C. Xue, Z.L. Zhang, H. Liu, Learning knowledge graph embedding with multi-granularity relational augmentation network, Expert Syst. Appl. 233 (2023) 120953.
[41] W.X. Wang, G.X. Zhang, P. Zhong, et al., Domain adaptation with contrastive and adversarial oriented transferable semantic augmentation, Knowl.-Based Syst. 282 (2023) 111092.
[42] J.J. Yin, Z.C. Zheng, Y.L. Pan, et al., Semi-supervised semantic segmentation with multi-reliability and multi-level feature augmentation, Expert Syst. Appl. 233 (2023) 120973.
[43] C.J. Si, Y.H. Jia, R. Wang, et al., Multi-label classification with high-rank and high-order label correlations, IEEE Trans. Knowl. Data Eng. 36 (8) (2024) 4076–4088.
[44] R.H. Shang, J.Y. Zhong, W.T. Zhang, et al., Multilabel feature selection via shared latent sublabel structure and simultaneous orthogonal basis clustering, IEEE Trans. Neural Netw. Learn. Syst. (2024) http://dx.doi.org/10.1109/TNNLS.2024.3382911.
[45] Z.F. Liu, C.J. Tang, S.E. Abhadiomhen, et al., Robust label and feature space co-learning for multi-label classification, IEEE Trans. Knowl. Data Eng. 35 (11) (2023) 11846–11859.
[46] J. Zhang, H.R. Wu, M. Jiang, et al., Group-preserving label-specific feature selection for multi-label learning, Expert Syst. Appl. 213 (2023) 118861.
[47] J.H. Ma, B.C.Y. Chiu, T.W.S. Chow, Multilabel classification with group-based mapping: A framework with local feature selection and local label correlation, IEEE Trans. Cybern. 52 (6) (2022) 4596–4610.
[48] R. Schapire, Y. Singer, A boosting-based system for text categorization, Mach. Learn. 39 (2/3) (2000) 135–168.
[49] J. Demsar, Statistical comparisons of classifiers over multiple data sets, J. Mach. Learn. Res. 7 (2006) 1–30.