International Journal of Machine Learning and Cybernetics (2023) 14:3737–3752
https://doi.org/10.1007/s13042-023-01861-2

ORIGINAL ARTICLE

Multi-granular labels with three-way decisions for multi-label classification

Tianna Zhao[1,2,3,4] · Yuanjian Zhang[5] · Duoqian Miao[1] · Hongyun Zhang[1]

Received: 30 September 2022 / Accepted: 11 May 2023 / Published online: 8 June 2023

Abstract

Multi-label classification is a challenging issue because it simultaneously embraces the characteristics of the imbalanced class distribution for each label and the uncertain label correlation among the whole label space. The decision-theoretic rough set can describe the roughness of concepts in the sense of minimizing decision risk but fails to consider the case where concepts are compatible. We argue that it is feasible to analyze the uncertainty of coarse-grained logical labels with limited label correlation assumptions and reduce the classification error for those uncertain instances by learning fine-grained numerical labels. Consequently, we develop a multi-granular label information system by introducing a multi-granular threshold with a three-way-based label enhancement (MGT-LEML) model. With the second-order label correlation assumption, we deduce the pseudo-positive and pseudo-negative classes for each label. The decision-theoretic rough set evaluates the possibility of misclassification independently, and a novel uncertain measure called instance uncertainty degree determines whether it is necessary to conduct label enhancement afterward. In this way, instances with the most uncertain classifications across label space compute fine-granule numerical labels by label enhancement, whereas remaining unchanged otherwise. We analyze the comparison results among nine algorithms on eight benchmarks with six metrics to demonstrate the superiority of the proposed MGT-LEML algorithm over state-of-the-art multi-label classification algorithms. Compared with the HNOML algorithm, our algorithm achieves significant improvement. Concretely, the performance is reduced by 2.9% in Hamming Loss, 12.4% in Ranking Loss, 14.3% in One Error, 465.5% in Coverage, and is increased by 14.2% in Average Precision.

Keywords Multi-label classification · Multi-granular · Three-way decisions · Decision-theoretic rough sets · Uncertainty · Label enhancement

1 Introduction

Label ambiguity is a widespread issue in multi-label learning [1–3]. The classifier learns a projection with known logical labels and determines the relevance between the instances and the labels for the unseen instances. It is applied widely across many practical applications, such as text semantic analysis [4], X-ray disease screening [5], crowdsourcing [6], scene classification [7], and age estimation [8, 9].

Most multi-label classification researches focus on the novelty of learning strategy. *Problem transformation* and *algorithm adaptation* are two representative strategies to design the multi-label classifier. The first group takes

T. Zhao and Y. Zhang contributed equally to this work.

✉ Duoqian Miao
dqmiao@tongji.edu.cn

Tianna Zhao
zhaotianna@shnu.edu.cn

Yuanjian Zhang
zhangyuanjian@unionpay.com

Hongyun Zhang
zhanghongyun@tongji.edu.cn

[1] Department of Computer Science and Technology, Tongji University, 4800 Cao'an Highway, Shanghai 201804, China

[2] Institute of Artificial Intelligence on Education, Shanghai Normal University, 100 Haisi Road, Shanghai 200234, China

[3] The Research Base of Online Education for Shanghai Middle and Primary Schools, Shanghai Normal University, 100 Haisi Road, Shanghai 200234, China

[4] Shanghai Engineering Research Center of Intelligent Education and Big data, Shanghai Normal University, 100 Haisi Road, Shanghai 200234, China

[5] China UnionPay Co. Ltd, 1699 Gutang Road, Shanghai 201201, China

advantage of the existing learning mechanism on single-label classification algorithms and promotes the multi-label model by decomposing it into a couple of subproblems. The processing on each subproblem is much similar to that of binary or multi-class cases. For this point, we have solutions like BR [10], RA*k*EL [11], LLSF [12] and TSEN [13]. In contrast, the latter group tailors the existing algorithms to satisfy the requirement of multi-output. Well-known works include ML*k*NN [14], ML-Forest [15], MLTSVM [16] and so on. Both learning strategies employ calibrated threshold to determine the final label associations [17]. For better generalization, the function of the classification model is mostly linear, and the calibrated threshold determines the position of the hyperplane. The instances around the hyperplane have a larger possibility of misclassification, and vice versa. Whatever the value of the calibrated threshold is, some instances are inevitably closer to hyperplane than others. Considering the multifaceted semantics and combinations of each label, the difficulties in constructing optimal classification are still challenging.

Conventional multi-label learning only concerns qualitative label relevance, but label distribution is a more generalized case and assumes that all labels can quantitatively describe some semantics of instances with varying description degrees. For an arbitrary instance, the description degree develops a data structure similar to a probability distribution, and learning with such supervised information is defined as label distribution learning [18]. With this assumption, a picture can have the descriptions like *vast sky*, *some seagull*, and *not similar to a boat*. Compared with the label space represented by logical labels (for comparison, the same picture may have logical labels as *sky*, *seagull* and *no boat*), and label distribution offers stronger supervisions [19, 20].

The acquisition of label distribution by manual annotation is costly as it requires sophisticated discrimination within similar instances. One alternative solution is to leverage label enhancement by learning numerical label representations based on smoothness assumptions from both the feature and label sides. Tao et al. [21] constructed a low-rank stacked matrix with vertically placed features and logical labels. Reconstruction constraints on both feature sides and label sides support the enhancement. Li et al. [22] generated label distribution by employing label propagation on fully-connected graphs over training instances. Xu et al. [23] devised a label recovery strategy by combining the fitness of instance-label reconstruction error with the assumption of neighbourhood-based label similarity. Shao et al. [24] extended the previous work by proposing a unified framework with numerical label regression and label enhancement. All the aforementioned papers advocated the superiority of label enhancement against logical labels for multi-label classification, yet they do not systematically discuss how to identify and rectify the performance degeneration from the most uncertain instances. In reality, this is an essential technique for people in searching for an effective and economic solution.

Granular computing emphasized the approximate formulation of information granules in analyzing uncertain concepts. Traditionally, the information granules are the deduction of expanded mathematical models from fuzzy sets, rough sets and so on [25–27]. While the enriched approximation semantics draw some elegant conclusions on the description and computation of concepts, the restricted portability originating from the model components becomes a barrier in dealing with complicated cases like multi-label classification.

Three-way decisions [28], originated from the granular computing method, are the classical methodologies in dealing with uncertainty. As theoretical research deepens, it becomes a theory of thinking in three, with the three subsequent procedures as trisecting, acting, and outcome (a.k.a. TAO [29–31]), respectively. The three-way decisions divide concepts into three parts that solve various practical problems, and the semantics of the three procedures are problem-dependent [32–37]. For a concept identification task, the trisecting procedure discriminates the instances into three different conditions, whereas the uncertain-driven third option serves for those uncertain instances. The acting procedure takes corresponding actions (acceptance, rejection, or further classification) in accordance with the trisecting, whereas the outcome procedure evaluates the performance of all certain decisions. For the three-way decisions on multi-label classifications [38–40], the three-way framework is either bounded by logical or numerical labels, which does not substantially boost the upper bound of multi-label classification.

This paper proposed a novel method called multi-granular labels with three-way decisions for multi-label classification (MGT-LEML). We argue that label enhancement and logical label-based models can be cooperated to address the multi-label classification. Given the label-specific separation margin, the model determines the label associations directly if instances are with a large separation margin across all labels, and reclassifies the associations with latent enhanced labels otherwise. Throughout the whole paper, we conclude the contributions into three aspects:

(1) We propose a novel information system with multi-granular labels. By combining the learned implicit numerical labels with explicit logical labels, we develop a novel formulation for minimizing classification uncertainty. This learning mechanism simulates the hierarchical perceptions of humans when facing uncertainty.

(2) Existing approaches improve classification effectiveness by either generating multi-granular features or finding an appropriate feature representation, but the learning target (i.e., labels) keeps the nature of single-granularity. In contrast, we leverage numerical labels and logical labels for uncertain instances and certain instances, respectively. It enriches three-way-based multi-label classification.

(3) The finer-granule numerical labels are only necessary for uncertain instances instead of the whole with the measurement of instance uncertainty degree. The instance uncertainty degree measures the uncertainty classification distribution over the label space and develops hierarchically. We optimize the local estimation of classification uncertainty by employing a decision-theoretic rough set.

We organize the remaining parts. Section 2 presents the preliminary; We propose the main idea for multi-label classification in Sect. 3; Sect. 4 designs the compared experiments and analyzes the experimental results; Sect. 5 discusses some open issues regarding the proposed method; Sect. 6 concludes the work.

2 Preliminaries

This section reviews some preliminaries regarding multi-label classification, label-specific feature learning and label enhancement, which will be components in proposed MGT-LEML.

2.1 Multi-label classification

Multi-label classification learns a projection from feature space to label space so that the associated labels for an unseen instance can be determined simultaneously. Label correlation is intensively studied to alleviate the imbalanced class distribution issue. Generally, there are three kinds of label correlation assumptions (i.e., first-order strategy, second-order strategy, and high-order strategy):

The first-order strategy is a straightforward extension of single-label classification in that it learns label association independently. ML*k*NN [14] determined the label association of instances by estimating the maximum posterior probability within the k-neighborhood. LIFT [41] improved the label association by enhancing the feature representation by calculating label-dependent clusters in kernel space. The second-order strategy takes a pairwise assumption of label correlation. LLSF [12] assumed the stronger the label correlation within two labels is, the more likely the features are shared. Glocal [42] exploited the global and local label correlations in latent label space to reduce the influence of missing labels. MDFS [43] analyzed the local label correlations in the feature manifold, which is then regularized by global label correlation. The high-order strategy takes the most considerations on label correlation. MASP [44] embedded the high-order label correlations for feature extraction and generates stable predictions for queried instance-label pair. For the sake of computation, some algorithms like fRA*k*EL [45], MLR [46], and AC*k*EL [47] construct a covering representation of label subset and achieve a balanced complexity between the subproblem count and the subproblem itself.

In this study, we consider second-order label correlation as they do not introduce much computational overhead and report acceptable accuracy.

2.2 Label-specific feature learning

Label-specific feature attempts to explore the characteristics of different labels by searching different feature combinations, which differs from finding an identical feature representation that works for the whole. The pioneering work, LIFT [41], generated label-specific features by employing k-means clustering. A fixed number of clusters from different perspectives characterize the underlying structure of positive and negative classes. With these augmented features, all kinds of binary classifiers complete the multi-label classification in the unit of the label. FRS-SS-LIFT [17] claimed that LIFT neglected the problems incurred from feature redundancy and label ambiguity, and alleviated the drawbacks by employing a fuzzy rough set. The rationality is two-fold: firstly, a fuzzy rough set is competitive in attribute reduction, which removes the irrelevant features before conducting a feature mapping; secondly, a fuzzy rough set presents a well-defined approximation for concept with fuzziness and roughness, and the removal of uncertain instances not only reduce loss in effectiveness but also brings in acceleration in efficiency. LLSF [12] argued that LIFT ignored the label correlations, and incorporated the second-order and high-order label correlation from observed label space. The second-order label correlation assumes that one label is at most correlated with another, and the high-order label correlation assumes that label correlation holds in different subsets of labels. This work extends to LSML [48], where a revised label-specific feature learning solves the degenerated performance incurred by missing labels. The label-specific feature learning classifier is trained with the recovered label to decrease the deviated estimation of label correlation. MULFE [49] constructed label-specific features by leveraging label correlation on feature mapping induced by LIFT. The adjustable weights on cluster centers embrace the maximal margin across labels. LSR-LSF [50] reviewed the sparseness of label-specific features and examined the label ambiguity by leveraging reshape operation on label space. With label propagation and cosine similarity constraint, the enriched label space exhibited more refined representations in numerical style while maintaining the second-order label correlation. SENCE [51] emphasized the degeneration of clustering randomness in LIFT and addressed it by employing the clustering ensemble technique. The mixture-based clustering ensemble adopted the expectation-maximization algorithm. However, it still fails to leverage label correlation.

In our work, we consider LLSF [12] as a component. The reasons are as follows:

1. It leverages the label correlation information, which is very important in boosting classification performance.
2. We assume that the logical labels are available across the entire label space in the training stage. The missing label issue is beyond our scope.
3. The primary goal is to explore the classification uncertainty induced by the logical label-based model. Therefore label enrichment on all labels is not preferred.
4. This algorithm is simple in its design and describes the label difference intuitively with different weights.

Based on the previous characteristics, the objective function for the i-th label is given as formula (1):

$$\min_{\mathbf{w}^i} \frac{1}{2}\left\|\mathbf{X}\mathbf{w}^i - \mathbf{y}^i\right\|_2^2 + \frac{\delta}{2}\sum_{j=1}^{l} r_{ij}\mathbf{w}^{i^\top}\mathbf{w}^j + \eta\left\|\mathbf{w}^i\right\|_1 \tag{1}$$

Where $\mathbf{X}$ denotes the features, $\mathbf{w}^i$ denotes the weights of features related to the label l_i, where the contribution of a particular feature is more significant if the corresponding weight is larger, and is irrelevant to l_i if the weight reaches zero. $\mathbf{y}^i$ denotes the ground-truth on label l_i. $r_{ij} = 1 - c_{ij}$, where c_{ij} quantifies the correlation strength between l_i and l_j calculated by cosine measure. The symbol δ and η are parameters. $\mathbf{w}^{i^\top}\mathbf{w}^j$ means the correlation of l_i and l_j from feature view. A higher correlation between label l_i and label l_j implies a larger inner product between $\mathbf{w}^i$ and $\mathbf{w}^j$, and vice versa. Regarding the objective functions with the alike of the form (1) as a whole, we rewrite the objective function as formula (2):

$$\min_{\mathbf{W}} \frac{1}{2}\|\mathbf{XW} - \mathbf{Y}\|_F^2 + \frac{\delta}{2}tr\left(\mathbf{R}\mathbf{W}^\top\mathbf{W}\right) + \eta\|\|\mathbf{W}\|\|_1 \tag{2}$$

Where $\mathbf{W} = \left[\mathbf{w}^1, \mathbf{w}^2, \dots, \mathbf{w}^l\right]$ denotes a weight matrix from all label-specific features in the order of label sequence. Label correlation matrix $\mathbf{R} = \left[r_{ij}\right]$ is calculated regarding second-order label relevance. The notation $tr(\cdot)$ denotes the rank of a matrix.

2.3 Label enhancement

Label enhancement attempts to reconstruct the latent but refined description between instances and labels by employing learning automatically instead of labeling artificially. The idea of label enhancement can be traced back to LEMLL [24]. This method leveraged topological information hidden on the feature side by employing local linear embedding, and the established optimization framework is in a wrapped manner. CFSM [52] explored a cost-sensitive feature reduction strategy as the pre-processing of label enhancement. By combining the neighbourhood-based granules for the feature side, label significance for logical labels, and the feature cost with representative probability distribution functions, a filtering-based feature selection criterion is defined. The features with higher scores served as refined input for further processing. BD-LE [53] claimed that information loss existed if only unidirectional projection from the feature side to the label is available. It mitigated this loss by developing a bidirectional loss function, where the inverse mapping from label side to feature offered the reconstruction error information. L^2 [54] presented an end-to-end solution involving both label enhancement and label distribution. In particular, an adaptive similarity graph constructed by locally linear embedding is alternatively optimized with label distribution learning. LELSF [55] argued that the projection from the feature to the enriched label may not be completely linear, and not all features contribute to the enrichment of an arbitrary label. In this work, the linear property hold from features with high dimensionality to numerical labels, where both label-specific and label-common feature components are involved. LEFND [56] alleviated the performance degeneration from redundant features by replenishing the fuzzy discrimination index. In this method, the label enhancement is independent of the label distribution learning, where label enhancement adopts some first-order statistics to estimate the soft connection between logical labels and numerical labels. MDLRML [57] exploited both feature manifold and label manifold based on smoothness assumption. Although the deduced multi-output regressor is competitive in enhancing local data fitness, the label enhancement module is still independent of the classifier training.

In our work, we consider LEMLL [24] as a component. The reasons are as follows:

1. The combination of label enhancement and logical label learning is an inspiring attempt to reduce uncertainty and contributes to knowledge representation.
2. It leverages adequate information reducing the ambiguity of label semantics (i.e., both topology information on the feature side and the closeness assumption on the label side) without introducing additional operations. The simplicity property seems more appropriate to explain the effectiveness of our idea.

Inspired by the previous knowledge, we have the following objective function:

Table 1 Notations of MGT-LEML

Notations	Mathematical meanings
$\lambda_{\cdot\cdot}$	loss function
D_l	multi-label instances set
D_u	unseen instances set
$\mathbf{X}$	feature space
$\mathbf{Y}$	label space
$\mathbf{x}_i$	an instance
$\mathbf{y}_i^1$	logical label
$\mathbf{y}_i^2$	numerical label
$\mathbf{x}_u$	an unseen instance
$\mathbf{x}_u^*$	an uncertainty instance in unseen instance set
$f^1(\cdot)$	logical label model learnt by LLSF
$f^2(\cdot)$	numerical label model learnt by LEMLL
$\mathbf{W}$	weight parameter in LLSF
$\boldsymbol{\phi}$	mapping from feauture to high dimension space
$\mathbf{b}, \theta$	parameter of linear mapping in LEMLL
$f_i^1\left(\mathbf{x}_u\right)$	the output on label l_i of logical label function learnt by LLSF
$f_i^2\left(\mathbf{x}_u^*\right)$	the output on label l_i of numerical label function learnt by LEMLL
$\mathbf{y}_u^*$	the label set learnt by f_2
$\hat{\mathbf{y}}_u$	the label set learnt by f_1
$D_{(\beta_0,\alpha_0)}$	uncertain instance set on l_i
$\neg D_{(\beta_0,\alpha_0)}$	certain instance set on l_i
$\mathbf{Y}_u$	final predicted multi-label set of $\mathbf{X}_2$

$$\begin{aligned}
&\min_{\boldsymbol{\Theta},\mathbf{b},\mathbf{U}} \sum_{i=1}^{n} L_R\left(R_i\right) + \mu\|\Theta\|_F^2 + \lambda\|\mathbf{U}-\mathbf{Y}\|_F^2 + \gamma tr\left(\mathbf{U}^\top\mathbf{M}\mathbf{U}\right)\\
&s.t. \quad R_i = \left\|\xi_i\right\|_2 = \sqrt{\xi_i^\top \xi_i};\\
&\qquad \boldsymbol{\xi}_i = \mathbf{u}_i - \boldsymbol{\Theta}\varphi\left(\mathbf{x}_i\right) - \mathbf{b}\\
&\qquad L_R(R) = \begin{cases} 0 & R < \varepsilon; \\ R^2 - 2R\varepsilon + \varepsilon^2 & R \geqslant \varepsilon. \end{cases}
\end{aligned} \tag{3}$$

Where we assume that the projection is linear, (i.e., $f(\mathbf{x}_i) = \boldsymbol{\Theta}\varphi\left(\mathbf{x}_i\right) + \mathbf{b}$), it maps the instances to a higher dimensional space and then does a linear mapping to predict label set. $\boldsymbol{\xi}_i = \mathbf{u}_i - \boldsymbol{\Theta}\varphi\left(\mathbf{x}_i\right) - \mathbf{b}$ is the difference between the restored numerical labels and the numerical labels predicted by the classifier and the loss function $\sum_{i=1}^{n} L_R\left(R_i\right)$ uses the idea of allowing for small errors, and R_i is the Euclidean distance, $\|\mathbf{U}-\mathbf{Y}\|_F^2$ constrains the restored numeric labels to be as similar as possible to the original logical labels. $tr\left(\mathbf{U}^\top\mathbf{M}\mathbf{U}\right) = \|\mathbf{U}-\mathbf{W}\mathbf{U}\|_F^2$ uses the smooth assumption that the organizational structure of the feature space is similar to that of the label space. $\mathbf{M} = (\mathbf{I}-\mathbf{W})^\top(\mathbf{I}-\mathbf{W})$, where $\mathbf{W}$ is the weight of graph $G = (\mathbf{V}, \mathbf{E}, \mathbf{W})$, and this component represents the associations on arbitrary two instances. In graph G, any instances within a distance $\boldsymbol{\xi}$ are ignored, whereas the remaining are approximately represented by the combinations of k-neighborhood, denoted as formula (4):

$$\begin{aligned}
&\min_{\mathbf{W}} \sum_{i=1}^{n} \left\| \mathbf{x}_i - \sum_{j \neq i} W_{ij}\mathbf{x}_j \right\|^2\\
&s.t. \quad \sum_{i=1}^{n} W_{ij} = 1.
\end{aligned} \tag{4}$$

3 The MLT-LEML model

3.1 Notations

Table 1 presents a nomenclature which elaborates on the major notations and the corresponding mathematical meanings.

3.2 Basic idea

MGT-LEML learns two projections (i.e., f^1 and f^2) sequentially from the feature $(\mathbf{X})$ to the label $(\mathbf{Y})$ and introduces two thresholds for the determination of logical and numerical labels in coarse and refined granules, respectively. Figure 1 illustrates the pipeline.

The learning target in our case is multi-granularity in the form of multi-label. Specifically, the coarse granularity is developed from the logical label model, whereas the fine granularity is constructed according to the requirement of

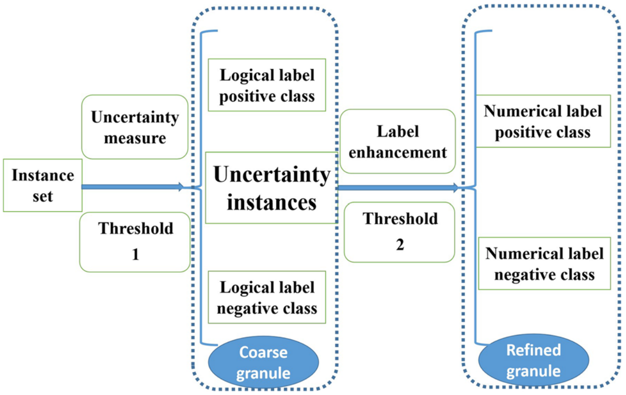

Fig. 1 Pipeline of MGT-LEML. The multi-granular thresholds serves for logical label learning (at coarse granule) and numerical label learning (at refined granule), respectively

uncertainty analysis. Then a novel information system is defined as Definition 1.

Definition 1 A multi-granular label information system (MGLIS) is a quadruple denoted as $MGLIS = (\mathbf{X}, \mathbf{Y}, V, f)$, where $\mathbf{X}$ is observed feature space and $\mathbf{Y}$ is the label space with multi-granularity; $f = \{f^1, f^2\}$ is the projection from the feature to label space, where f^1 and f^2 identifies the coarse and fine label, respectively. The coarse label of instance $\mathbf{x}_i$ is represented by $\mathbf{y}_i^1$, which satisfies $\mathbf{y}_i^1 \in \{0, 1\}^l$. In contrast, the fine label of instance $\mathbf{x}_i$ is represented by $\mathbf{y}_i^2$, which satisfies $\mathbf{y}_i^2 \in [-1, 1]^l$ if available and NA otherwise.

Definition 1 shows that an instance may have at most two levels of the label, where the coarse level and refined label correspond to the logical and numerical label, respectively. To facilitate understanding, we present an example of MGLIS in Table 2.

In what follows, we elaborate on how to establish and apply the MGLIS. Concretely, we discuss the functionality of multi-granular thresholds (see Sect. 3.3) and how they work for multi-label classification (see Sect. 3.4).

3.3 Multi-granular label representations

The classification for coarse granule label representation (i.e., the logical labels) follows the settings of LLSF [12] such that the label-specific features determine the logical labels.

However, the classification for fine granule label representation (i.e., the numerical labels) has some differences to LEMLL [24] in that the instances to be enhanced are those uncertain instances instead of the whole. The rationality is that given linear projection learned by f^1, instances with smaller separation margins to hyperplane have a larger possibility of misclassification, and vice versa. In other words, the classification results have a smaller possibility for corrections if we conduct label enhancement on the instances of certain classifications. Consequently, the refined-granule labels in Table 2 are not invariably available (see instance $\mathbf{x}_7$ and $\mathbf{x}_8$).

Different from the single-label case in that the separation margin corresponds to only linear projection, the separation margins exhibit a collection of margins for each label. This means we can devise different strategies to find uncertain instances. For those uncertain instances $\mathbf{x}_u^* \in D_{(\beta,\alpha)}$, f^2 generates the label distribution and completes the classification with a new virtual label u_0. Therefore, the final predicted label $\mathbf{y}_u \in \mathbf{Y}_u$ is shown as formula (5):

$$\mathbf{y}_u = \begin{cases} \hat{\mathbf{y}}_u, & \mathbf{x}_u \in \neg D_{(\beta_0,\alpha_0)}; \\ \mathbf{y}_u^*, & \mathbf{x}_u \in D_{(\beta_0,\alpha_0)}. \end{cases} \tag{5}$$

Table 2 An illustration of multi-granular label representation

Instances	Attributes			Coarse granule			Refined granule		
				Logical labels			Numerical labels		
	a_1	a_2	a_3	l_1	l_2	l_3	u_1	u_2	u_3
$\mathbf{x}_1$	1	1	0	1	0	1	0.4	−0.2	0.4
$\mathbf{x}_2$	1	0	0	1	0	1	0.6	−0.1	0.3
$\mathbf{x}_3$	1	1	0	1	1	0	0.5	0.4	−0.1
$\mathbf{x}_4$	1	0	1	1	1	0	0.6	0.3	−0.1
$\mathbf{x}_5$	0	1	1	0	1	1	−0.1	0.5	0.4
$\mathbf{x}_6$	0	1	1	0	1	1	−0.1	0.3	0.6
$\mathbf{x}_7$	0	0	0	1	1	0	NA	NA	NA
$\mathbf{x}_8$	1	1	1	0	1	1	NA	NA	NA

3.4 Label enhancement with three-way decisions

For an unseen instance $\mathbf{x}_u \in D_u$, we specify which label predictions are uncertain by computing regression based on the formula (6).

$$f_1(\mathbf{x}_u) = \mathbf{W}\mathbf{x}_u. \tag{6}$$

Where $\mathbf{W} = [\mathbf{w}^1, \mathbf{w}^2, \dots, \mathbf{w}^l]$ is learnt from formula (1). Due to the linear model property, the calibrated threshold l_0 constitutes a hyperplane with $f_1^i(\mathbf{x}_.) = l_0, \forall i \in \{1, 2, \dots, l\}$. Instances whose regression values are close to l_0 are more uncertain than those with a larger separation margin to the l_0. In what follows, we will show how to pick the uncertainty instances on each label.

The associated labels in multi-label are sparsely distributed across the label space, leading to even more imbalanced class distributions for each label. This means the count of the associated label (i.e., positive class, denoted by 1) can be far smaller than the count of the unassociated label (i.e., negative class, denoted by 0). In this case, we require an uncertain theory that embraces the following properties:

1. Cost-sensitive: It should indicate the cost difference for two representative misclassifications. That is, the decision cost of judging a label that should be relevant to an example as irrelevant should be significantly higher than the decision cost of deciding a label that should be irrelevant to an example to be relevant.
2. Boundary awareness: It should allow some vagueness for labels to seek the appropriate actions for deferment decisions when the confidence of evidence is not adequate.

Fortunately, the decision-theoretic rough set (DTRS) [28] satisfies the above-mentioned requirements. For a classical binary classification problem, it enumerates all possible combinations (six states) given three actions (acceptance, rejection and deferment) and presents a set of loss functions (see Table 3, denoted as $\lambda_{..}$). Meanwhile, the decision cost incurred by taking different actions is determined by the cumulative sum of the combination of the loss function with evidence measured by conditional probability.

The semantics of detailed settings of the six loss functions are as follows.

Table 3 Loss Functions for Decisions with Two States

	a_p	a_b	a_n
$[\mathbf{x}]_{l_i}$	λ_{ap}	λ_{bp}	λ_{np}
$\neg[\mathbf{x}]_{l_i}$	λ_{an}	λ_{bn}	λ_{nn}

1. When an instance is actually with a label l_i and we also decide that the instance has the label l_i, then the decision risk λ_{ap} should be very low.
2. When an instance is actually with a label l_i but we do not decide the label-instance association, then the decision risk λ_{bp} should be higher than that of λ_{ap}.
3. When an instance is actually with a label l_i but we decide the instance is without the label l_i, then the decision risk λ_{np} should be very high.
4. When an instance is actually without a label l_i and we also decide that the instance is without label l_i, then the decision risk λ_{nn} should be very low.
5. When the instance is actually without a label l_i but we do not decide the label-instance association, the decision risk λ_{bn} should be higher than λ_{nn}.
6. When an instance is actually without a label l_i but we decide that the instance is with the label l_i, then the decision risk λ_{an} should be very high.

The six losses cover all possible actions when taking three-way decisions.

Generally, the following two assumptions are satisfied:

(1) $0 \leqslant \lambda_{ap} \leqslant \lambda_{bp} \leqslant \lambda_{np}, 0 \leqslant \lambda_{nn} \leqslant \lambda_{bn} \leqslant \lambda_{an}$;

(2) $\frac{\lambda_{np}-\lambda_{bp}}{\lambda_{bn}-\lambda_{nn}} > \frac{\lambda_{bp}-\lambda_{ap}}{\lambda_{an}-\lambda_{bn}}$.

We denote $P([\mathbf{x}]_{l_i}|\mathbf{x})$ as the probability that an instance $\mathbf{x}$ is with label l_i condition on the instances $\mathbf{x}$. The decision risks that instance is with label l_i (acceptance), defer to the decision (deferment) is without the label l_i (rejection), as shown in formula (7):

$$\begin{aligned} Cost(a_p|\mathbf{x}) &= \lambda_{ap}P([\mathbf{x}]_{l_i}|\mathbf{x}) + \lambda_{an}P(\neg[\mathbf{x}]_{l_i}|\mathbf{x}); \\ Cost(a_b|\mathbf{x}) &= \lambda_{bp}P([\mathbf{x}]_{l_i}|\mathbf{x}) + \lambda_{bn}P(\neg[\mathbf{x}]_{l_i}|\mathbf{x}); \\ Cost(a_n|\mathbf{x}) &= \lambda_{np}P([\mathbf{x}]_{l_i}|\mathbf{x}) + \lambda_{nn}P(\neg[\mathbf{x}]_{l_i}|\mathbf{x}). \end{aligned} \tag{7}$$

Based on the Bayesian minimizing decision cost principle, the decision risk $Cost(a_p|\mathbf{x})$ indicating the instance $\mathbf{x}$ is with label l_i (acceptance) reaches the minimum if

$$Cost(a_p|\mathbf{x}) \leqslant Cost(a_b|\mathbf{x}). \tag{8}$$

$$Cost(a_p|\mathbf{x}) \leqslant Cost(a_n|\mathbf{x}). \tag{9}$$

hold.

The decision risk $Cost(a_b|\mathbf{x})$ do not decide label association between the instance $\mathbf{x}$ and the label l_i (deferment) reaches the minimum, if

$$Cost\left(a_b|\mathbf{x}\right) \leqslant Cost\left(a_p|\mathbf{x}\right). \tag{10}$$

$$Cost\left(a_b|\mathbf{x}\right) \leqslant Cost\left(a_n|\mathbf{x}\right). \tag{11}$$

hold.

The decision risks $Cost\left(a_n|\mathbf{x}\right)$ indicating that instance $\mathbf{x}$ is without the label l_i (rejection) reaches the minimum if

$$Cost\left(a_n|\mathbf{x}\right) \leqslant Cost\left(a_p|\mathbf{x}\right). \tag{12}$$

$$Cost\left(a_n|\mathbf{x}\right) \leqslant Cost\left(a_b|\mathbf{x}\right). \tag{13}$$

hold.

β and α can be computed as formula (14) and formula (15), respectively.

$$\beta = \frac{\lambda_{bn} - \lambda_{nn}}{\left(\lambda_{bn} - \lambda_{nn}\right) + \left(\lambda_{np} - \lambda_{bp}\right)}. \tag{14}$$

$$\alpha = \frac{\lambda_{an} - \lambda_{bn}}{\left(\lambda_{an} - \lambda_{bn}\right) + \left(\lambda_{bp} - \lambda_{ap}\right)}. \tag{15}$$

For taking actions, we accept $\mathbf{x} \in [\mathbf{x}]_{l_i}$ if $P\left([\mathbf{x}]_{l_i}|\mathbf{x}\right) \geqslant \alpha$, reject $\mathbf{x} \in [\mathbf{x}]_{l_i}$ if $P\left([\mathbf{x}]_{l_i}|\mathbf{x}\right) \leqslant \beta$ and defer to decide $\mathbf{x} \in [\mathbf{x}]_{l_i}$ otherwise.

Due to the imbalanced class distribution of multi-label, the paper introduces the same six loss functions on all labels independently and deduces the trisecting based on formula (14) and (15). Therefore, the classification on label l_i is uncertain if $P\left([\mathbf{x}]_{l_i}|\mathbf{x}_u\right) \in (\beta, \alpha)$.

Recall that the output of LLSF on an arbitrary label l_i (that is $f_1^i(\mathbf{x}_u)$) satisfies $f_1^i(\mathbf{x}_u) \in (-0.5, 1.5)$, we can approximately regard the value $f_1^i(\mathbf{x}_u)$ as the conditional probability of $\mathbf{x}_u$ being the positive label (that is $P([\mathbf{x}]_{l_i}|\mathbf{x}_u)$). The tri-partition threshold (i.e., $\left(\beta_0, \alpha_0\right)$) in the sense of original regression can be estimated as:

$$\beta_0 = 2\beta - 0.5. \tag{16}$$

$$\alpha_0 = 2\alpha - 0.5. \tag{17}$$

Based on Bayesian minimizing cost principle, we deduce the three-way classification as formula (18):

$$f_1^i\left(\mathbf{x}_u\right) = \begin{cases} 1 & f_1^i\left(\mathbf{x}_u\right) \geqslant \alpha_0; \\ 0.5 & \beta_0 < f_1^i\left(\mathbf{x}_u\right) < \alpha_0; \\ 0 & f_1^i\left(\mathbf{x}_u\right) \leqslant \beta_0. \end{cases} \tag{18}$$

Where $0 \leqslant \beta_0 < l_0 < \alpha_0 \leqslant 1$ and $1 \leqslant i \leqslant l$.

Unlike single-label which the uncertainty of instances is equivalent to the uncertainty of labels, it is likely for instances with some uncertain labels. Formally, let

$$\hat{\mathbf{y}}_u = \left\{f_1^1\left(\mathbf{x}_u\right), f_1^2\left(\mathbf{x}_u\right), \ldots, f_1^l\left(\mathbf{x}_u\right)\right\}. \tag{19}$$

denote the multi-label classification on instance $\mathbf{x}_u$, $f_1^i\left(\mathbf{x}_u\right) \geqslant \alpha_0$ and $f_1^j\left(\mathbf{x}_u\right) \leqslant \beta_0$ may hold simultaneously. To correct the possible misclassifications, we take the optimistic strategy and define the uncertain instances selection principle $D_{(\beta_0,\alpha_0)}$ as formula (20):

$$D_{(\beta_0,\alpha_0)} = \cup\left\{\mathbf{x}_u | \exists i \in \{1, 2, \ldots, l\} \wedge \beta_0 < f_1^i\left(\mathbf{x}_u\right) < \alpha_0\right\}. \tag{20}$$

It is worth mentioning that the $D_{(\beta_0,\alpha_0)}$ is a realization of the acting procedure in three-way decisions. An instance is uncertain if it contains at least an uncertain label prediction and defers the classification until we obtain the latent label distributions. For an undetermined instance $\mathbf{x}_u^* \in D_{(\beta_0,\alpha_0)}$, the regression estimation of label distribution forms into the formula (21).

$$f_2\left(\mathbf{x}_u^*\right) = \mathbf{\Theta}\varphi\left(\mathbf{x}_u^*\right) + \mathbf{b}. \tag{21}$$

For each label l_i, we can determine the label relevance regarding $\mathbf{x}_u^*$ by introducing a virtual label u_0, as described in formula (22).

$$f_2^i\left(\mathbf{x}_u^*\right) = \begin{cases} 1 & f_2^i\left(\mathbf{x}_u^*\right) \geqslant u_0; \\ 0 & f_2^i\left(\mathbf{x}_u^*\right) < u_0. \end{cases} \tag{22}$$

Formally, let

$$\mathbf{y}_u{}^* = \left\{f_2^1\left(\mathbf{x}_u^*\right), f_2^2\left(\mathbf{x}_u^*\right), \ldots, f_2^l\left(\mathbf{x}_u^*\right)\right\}. \tag{23}$$

denote the multi-label classification on instance $\mathbf{x}_u^*$.

3.5 Algorithm complexity

We introduce multi-granular labels with three-way decisions for multi-label classification in Algorithm 1.

Algorithm 1 MGT-LEML

```
Require: Known labeling instances $D_l$, balance factors $\delta$, $\eta$, $\mu$, $\lambda$, $\gamma$, $\varepsilon$.
Ensure: $\mathbf{y}_u$ on unseen instances $\mathbf{Y}_u$.
1:  Generate $\mathbf{W}$ based on objective function defined in formula (2) based on $D_l$, $\delta$, and $\eta$.
2:  Generate $\mathbf{\Theta}$, $\mathbf{b}$ based on objective function defined in formula (3) based on $D_l$, $\mu$, $\lambda$, $\gamma$, $\varepsilon$.
3:  Generate $\beta$ and $\alpha$ based on formula (14) and (15).
4:  Generate $\beta_0$ and $\alpha_0$ based on formula (16) and (17).
5:  Generate regression results regarding LLSF ($f_1(\mathbf{x})$) based on $D_u$ and formula (6).
6:  for $i = 1$ to $l$ do
7:      Compute $f_1^i(\mathbf{x}_u)$ as described in formula (18).
8:  end for
9:  Generate $\hat{\mathbf{y}_u}$ as described in formula (19).
10: Generate $D_{(\beta_0,\alpha_0)}$ as described in formula (20).
11: for $u = 1$ to $|D_u|$ do
12:     for $i = 1$ to $l$ do
13:         if $\mathbf{x}_\mathbf{u}^* \in D_{(\beta_0,\alpha_0)}$ then
14:             Compute $f_2\left(\mathbf{x}_u^*\right) \triangleq \mathbf{y}_u^*$ as described in formula (21).
15:             Determine $f_2^i\left(\mathbf{x}_u^*\right) \in \{0,1\}^l$ as described in formula (22).
16:         end if
17:     end for
18:     Generate $\mathbf{y}_u^*$ as described in formula (23).
19: end for
20: Generate $\mathbf{y}_u \in \mathbf{Y}_u$ as described in formula (5).
```

The quantitative analysis of the complexity of Algorithm 1 explains step by step. Step 1 occupies $O\left(t_1\left(d^2 l + dl^2\right) + d^3 + l^3\right)$, where t_1 is the iteration count for solving LLSF, d is the count of feature dimensionality, and l is the count of labels. Step 2 occupies $O\left(t_2|D_l|^2\left(d^2 l + dl^2\right) + d^3\right)$, where t_2 is the iteration count for solving LEMLL. Step 3 and step 4 occupy the $O(1)$ for threshold generation. The complexity from step 5 to step 8 is $O\left(|D_u|l\right)$. Step 9 costs $O(l)$, whereas step 10 takes $O\left(|D_u|l\right)$. From step 11 to step 19, the computational complexity is $O\left(|D_{(\beta_0,\alpha_0)}|l\right)$. Step 20 takes the complexity of $O\left(|D_{(\beta_0,\alpha_0)}|l\right)$. Note that $t_2 \gg t_1$, $|D_{(\beta_0,\alpha_0)}| \ll |D_l|$, thus the overall the complexity of MGT-LEML is $O\left(t_2|D_l|^2\left(d^2 l + dl^2\right) + d^3\right)$.

Table 4 Data characteristics

Data set	# Instances	# Features	# Labels	# Cardinality	Domain
bibtex	7395	1836	159	2.402	text
birds	645	260	19	1.014	audio
emotions	593	72	6	1.869	music
enron	1702	1001	53	3.39	text
genbase	662	1185	27	1.252	biology
medical	978	1449	45	1.245	text
langualgelog	1460	1004	75	1.18	text
scene	2407	294	6	1.074	image

4 Experiments

4.1 Settings

We evaluate the classification effectiveness of MGT-LEML on eight publicly available benchmarks[58]. These datasets are of small or medium size, as the complexity of MGT-LEML is considerable. Table 4 describes the characteristics of the considered datasets, including the instances count (# Instance), the features count (# Features), the labels count (# Labels), the average count of associated labels per instances (# Cardinality) and the corresponding domains.

Table 5 Loss functions settings for decisions of label l_i

	a_p	a_b	a_n
$y_{*i} = 1$	λ_{ap} : 0	λ_{bp} : $3v$	λ_{np} : $8v$
$y_{*i} = 0$	λ_{an} : $7v$	λ_{bn} : $2v$	λ_{nn} : 0

We examine whether MGT-LEML gains superior classification performance against eight logical label-based algorithms. The comparisons of MGT-LEML include ML*k*NN, LIFT, MLTSVM, Glocal, HNOML, fRA*k*EL, and MCGM. Next, we will introduce the detail.

- ML*k*NN[1] [14]: It is an algorithm adaption of *k*NN that generates the multi-output simultaneously. The parameter *k* sets 10.
- LIFT[2] [41]: It learns label-specific feature mapping based on *k*-means clustering. The ratio parameter sets among $\{0.1, 0.2, \dots, 0.5\}$.
- LLSF[3] [12]: It learns label-specific features while preserving second-order label correlation. δ, η set among $\{2^{-10}, 2^{-9}, \dots, 2^9, 2^{10}\}$. The parameter τ_1 sets 0.5.
- MLTSVM[4] [16]: It learns distance differences in kernel space from multiple non-parallel hyperplanes. The penalty coefficient and kernel parameter are searched in $\{2^{-6}, 2^{-5}, \dots, 2^5, 2^6\}$ and $\{2^{-4}, 2^{-3}, \dots, 2^3, 2^4\}$, respectively.
- Glocal[5] [42]: It learns a function from explicit feature space to latent labels while capturing both global and local label correlation in the form of second-order. The penalty parameter sets 1.
- HNOML [59]: It leverages the data locality by imposing label embedding and label enriching. Penalty parameters α, β, γ set among $\{10^{-3}, 10^{-2}, \dots, 10^3\}$.
- fRA*k*EL[6] [45]: It selects a proportion of key instances to conduct Random *k*-label sets. The label set size takes the empirical value of 3, whereas the base classifier count is configured as twice the label cardinality count. The base classifier adopts LibLinear.[7]
- MCGM[8] [60]: It learns group-based projection by leveraging group-based local correlation with label-specific but local features. The parameter λ and β are searched in $\{10^{-4}, 10^{-3}, \dots, 1\}$, whereas α and δ are searched in $\{10^{-2}, 10^{-1}, \dots, 10^2\}$.
- MGT-LEML: Proposed method. f_1 takes the following settings ($\delta = 2^8, \eta = 2^4$, maximal iteration count t_1 is 100 and calibrated threshold $l_0 = 0.5$), whereas f_2 takes the following settings ($k = 10, \varepsilon = 0.1, \mu = \frac{1}{4}, \lambda = 1, \gamma = \frac{1}{64}$, calibrated threshold $u_0 = 0$, maximal iteration count t_2 is 200, $\phi(\cdot)$ adopts linear kernel function). For simplicity, the six losses for each label adopt the recommendation in [38], as shown in Table 5, where $y_{*i} = 1$ represents the event that instances are associated with label l_i, and otherwise if $y_{*i} = 0$.

Based on formula (14) and (15), the pair of thresholds β and α are computed respectively as $\frac{2}{7}$ and $\frac{5}{8}$. The adjusted thresholds β_0 and α_0 are $\frac{1}{14}$ and $\frac{3}{4}$ by applying formula (16) and (17), respectively.

We use six metrics [61] (*Hamming Loss*, *Ranking Loss*, *One Error*, *Coverage*, *Average Precision*, and *Micro F1*) to measure classification performance. The first four metrics report better performance if the values are smaller, denoted as the notation ↓. For metrics *Average Precision* and *Micro F1*, the larger the values are, the better the performance will be, denoted as ↑. The value labelled in brackets ranks the algorithm performance, and Avg rank shows the average ranking list of each algorithm on all data sets. We annotate the best performance value in bold size. All experiments are implemented by Matlab on a desktop with Intel(R) Core(TM) i7 processor and 32GB RAM (Table 6).

4.2 Results

Table 6 enumerates the representative statistics of classification performance measured by the six evaluation metrics from the nine algorithms. The average performance on considered metrics is from six times five-fold cross-validation and is independent of data randomness. Based on the reported average performance, we rank the algorithms as the metric declares from the best to the worst. The best performance is in bold size. From the metric view, MGT-LEML ranks first at 66.67% and second at 16.67% 16.67%. From the dataset view, MGT-LEML ranks first at 47.92% ($\frac{23}{48}$), second at 20.83% ($\frac{10}{48}$) and third at 6.25% ($\frac{3}{48}$). It achieves the least loss on metric *Coverage* (with 100% in the first place), whereas becomes the worst on metric *Micro F1* (with an average ranking in fourth place). Meanwhile, we observe that the performance of MGT-LEML improves at 87.5% ($\frac{42}{48}$) compared with LLSF, which further validates the effectiveness of MGT-LEML.

[1] code available at http://www.lamda.nju.edu.cn/code_MLkNN.ashx

[2] code available at http://cse.seu.edu.cn/PersonalPage/zhangml/index.htm

[3] code available at https://jiunhwang.github.io/

[4] code available at http://www.optimal-group.org/Resource/MLTSVM.html

[5] code available at http://www.lamda.nju.edu.cn/code_Glocal.ashx

[6] code available at http://github.com/KKimura360/fast_RAkEL_matlab

[7] code available at https://www.csie.ntu.edu.tw/~cjlin/liblinear/

[8] code available at https://github.com/JianghongMA/MC-GM

Table 6 Comparisons of algorithms on first-order statistics mean and second-order statistics standard deviation (separated by ±)

Data set	MGT-LEML	ML*k*NN	LIFT	LLSF	MLTSVM	Glocal	HNOML	fRA*k*EL	MCGM
	Hamming Loss (↓)								
bibtex	**0.012±0.001**(1)	0.016±0.001(7)	0.013±0.002(3)	0.013±0.001(3)	0.017±0.001(8)	0.014±0.001(5)	0.015±0.001(6)	0.013±0.001(3)	0.047±0.006(9)
birds	**0.051±0.002**(1)	0.056±0.004(4)	0.054±0.003(3)	0.094±0.006(7)	0.099±0.004(8)	0.053±0.005(2)	0.065±0.007(5)	0.185±0.009(9)	0.067±0.005(6)
emotions	**0.200±0.014**(1)	0.286±0.017(6)	0.254±0.016(5)	0.236±0.020(3)	0.287±0.016(7)	0.311±0.004(8)	0.230±0.014(2)	0.424±0.010(9)	0.248±0.033(4)
enron	0.046±0.002(3)	0.051±0.002(5)	0.046±0.001(3)	0.067±0.002(8)	0.062±0.002(6.5)	0.076±0.006(9)	0.046±0.001(3)	**0.045±0.001**(1)	0.062 ± 0.004(6.5)
genbase	**0.001±0.001**(2)	0.004±0.001(9)	0.002±0.001(4.5)	**0.001±0.001**(2)	0.002±0.001(4.5)	0.003±0.003(7)	0.003±0.003(7)	**0.001±0.001**(2)	0.003 ± 0.001(7)
languagelog	**0.015±0.001**(1)	0.016±0.001(3)	0.029±0.001(9)	0.025±0.001(6.5)	0.018±0.001(5)	0.016±0.001(3)	0.016±0.001(3)	0.025±0.001(6.5)	0.028 ± 0.001(8)
medical	**0.011±0.001**(1)	0.015±0.001(6)	0.013±0.001(2)	0.016±0.001(7.5)	0.014±0.002(4)	0.019±0.003(9)	0.014±0.042(4)	0.014±0.001(4)	0.016 ± 0.001(7.5)
scene	0.102±0.005(4)	0.092±0.006(2)	**0.079±0.005**(1)	0.113±0.006(6)	0.143±0.003(8)	0.111±0.006(5)	0.147±0.011(9)	0.093±0.004(3)	0.119 ± 0.005(7)
Avg rank	1.7500(1)	5.2500(5)	3.8125(2)	5.3750(6)	6.3750(8)	6.0000(7)	4.8750(4)	4.6875(3)	6.8750(9)
	Ranking Loss (↓)								
bibtex	0.082±0.004(2)	0.205±0.004(6)	**0.074±0.004**(1)	0.124±0.005(3)	0.660±0.006(9)	0.160±0.004(5)	0.139±0.005(4)	0.233±0.010(7)	0.252±0.012(8)
birds	**0.099±0.013**(1)	0.167±0.014(2)	N/A(9)	0.581±0.036(8)	0.292±0.015(6)	0.272±0.019(4)	0.205±0.033(3)	0.302±0.028(7)	0.281±0.035(5)
emotions	**0.166±0.021**(1)	0.270±0.007(5)	0.233±0.017(3)	0.446±0.028(9)	0.295±0.032(6)	0.402±0.007(8)	0.241±0.019(4)	0.401±0.028(7)	0.212±0.022(2)
enron	0.092±0.003(2.5)	0.092±0.003(2.5)	**0.077±0.006**(1)	0.188±0.007(5)	0.499±0.012(7)	0.132±0.008(4)	0.667±0.014(8)	0.679±0.019(9)	0.404 ± 0.014(6)
genbase	**0.001±0.000**(1.5)	0.005±0.005(7)	0.004±0.001(6)	**0.001±0.001**(1.5)	0.006±0.008(8)	0.002±0.001(4)	0.002±0.002(4)	0.002±0.001(4)	0.031 ± 0.037(9)
languagelog	0.147±0.015(2)	**0.127±0.005**(1)	0.150±0.013(3)	0.228±0.020(5)	0.731±0.014(9)	0.193±0.006(4)	0.288±0.018(6)	0.555±0.020(8)	0.311 ± 0.021(7)
medical	**0.016±0.007**(1)	0.043±0.007(5)	0.029±0.006(3)	0.051±0.012(6)	0.168±0.018(7)	0.038±0.006(4)	0.021±0.059(2)	0.206±0.022(8)	0.227 ± 0.023(9)
scene	0.080±0.013(2)	0.085±0.008(3)	**0.064±0.007**(1)	0.113±0.011(7)	0.278±0.009(9)	0.096±0.005(5)	0.110±0.010(6)	0.155±0.024(8)	0.095 ± 0.010(4)
Avg rank	1.6250(1)	3.9375(3)	3.3750(2)	5.5625(6)	7.625(9)	4.7500(5)	4.6250(4)	7.2500(8)	6.2500(7)
	One Error (↓)								
bibtex	0.344±0.020(2)	0.588±0.008(7)	0.386±0.010(4)	0.379±0.017(3)	0.424±0.013(5)	0.525±0.135(6)	0.618±0.005(8)	**0.195±0.008**(1)	0.752 ± 0.025(9)
birds	0.657±0.024(4)	0.854±0.036(8)	0.856±0.013(9)	**0.359±0.051**(1)	0.679±0.056(6)	0.690±0.036(7)	0.659±0.021(5)	0.563±0.026(3)	0.553±0.030(2)
emotions	0.255±0.028(4)	0.450±0.031(7)	0.366±0.022(6)	0.165±0.030(2)	0.470±0.025(8)	0.554±0.021(9)	0.253±0.038(3)	**0.157±0.030**(1)	0.333±0.136(5)
enron	0.217±0.010(2)	0.255±0.019(4)	0.240±0.021(3)	0.366±0.030(6)	**0.139±0.007**(1)	0.293±0.037(5)	0.950±0.017(8)	0.955±0.021(9)	0.760 ± 0.033(7)
genbase	0.001±0.003(3.5)	0.012±0.009(8)	**0.000±0.000**(1.5)	0.002±0.003(5.5)	0.001±0.001(3.5)	0.002±0.003(5.5)	0.006±0.003(7)	**0.000±0.000**(1.5)	0.144 ± 0.060(9)
languagelog	0.712±0.018(4)	0.722±0.015(5)	0.676±0.013(2)	0.804±0.018(7)	0.707±0.026(3)	0.983±0.011(9)	0.773±0.009(6)	**0.401±0.028**(1)	0.852 ± 0.036(8)
medical	0.150±0.018(3)	0.235±0.010(8)	0.165±0.014(6)	0.219±0.020(7)	0.113±0.020(2)	0.157±0.020(5)	0.154±0.177(4)	**0.062±0.016**(1)	0.510 ± 0.059(9)
scene	0.224±0.024(5)	0.243±0.171(6)	0.194±0.021(3)	0.288±0.027(8)	0.177±0.013(2)	0.264±0.011(7)	0.292±0.014(9)	**0.061±0.015**(1)	0.208 ± 0.160(4)
Avg rank	3.4375(2)	6.6250(7.5)	4.3125(4)	4.9375(5)	3.8125(3)	6.6875(9)	6.2500(6)	2.3125(1)	6.6250(7.5)
Data set	MGT-LEML	ML*k*NN	LIFT	LLSF	MLTSVM	Glocal	HNOML	fRA*k*EL	MCGM
	Coverage (↓)								
bibtex	**0.001±0.000**(1)	0.333±0.004(4)	22.14±1.176(6)	0.002±0.001(2)	0.503±0.005(5)	37.60±1.113(8)	0.211±0.006(3)	25.52±1.356(7)	1459 ± 32.58(9)
birds	**0.007±0.001**(1)	0.190±0.016(5)	3.752±0.350(6)	4.364±0.270(7)	7.532±0.298(8)	0.174±0.013(3)	0.009±0.001(2)	0.183±0.029(4)	109.6±12.45(9)
emotions	**0.051±0.005**(1)	0.383±0.006(4)	2.158±0.171(6)	2.590±0.119(8)	2.461±0.226(7)	0.496±0.009(5)	0.055±0.003(2)	0.359±0.033(3)	122.0±9.094(9)
enron	**0.005±0.002**(1)	0.246±0.008(3)	11.94±0.721(6)	0.008±0.001(2)	30.89±0.774(8)	18.04±0.872(7)	0.831±0.019(4)	0.835±0.040(5)	328.3 ± 9.465(9)
genbase	**0.000±0.001**(1)	0.212±0.008(3)	0.408±0.054(8)	0.001±0.001(2)	0.292±0.167(4)	0.352±0.081(6)	0.299±0.106(5)	0.372±0.166(7)	11.37 ± 5.464(9)
languagelog	**0.003±0.000**(1)	0.159±0.008(3)	13.39±1.557(4)	0.004±0.001(2)	30.79±1.403(8)	17.86±0.308(6)	27.56±1.346(7)	14.49±0.848(5)	163.4 ± 4.265(9)

Table 6 (continued)

Data set	MGT-LEML	ML*k*NN	LIFT	LLSF	MLTSVM	Glocal	HNOML	fRA*k*EL	MCGM
medical	**0.001±0.000**(1)	0.059±0.007(3)	1.954±0.342(5)	0.002±0.001(2)	4.648±0.700(7)	2.397±0.286(6)	8.123±1.268(8)	1.697±0.426(4)	78.17 ± 10.64(9)
scene	**0.013±0.001**(1)	0.085±0.006(3)	0.395±0.041(5)	0.018±0.001(2)	0.887±0.052(8)	0.568±0.033(7)	0.230±0.065(4)	0.551±0.050(6)	444.7 ± 34.90(9)
Avg rank	1.0000(1)	3.5000(3)	5.7500(6)	3.3750(2)	6.8750(8)	6.0000(7)	4.3750(4)	5.1250(5)	9.0000(9)
	Average Precision (↑)								
bibtex	**0.607±0.013**(1)	0.360±0.007(6)	0.561±0.006(2)	0.552±0.011(3)	0.326±0.012(8)	0.358±0.008(7)	0.362±0.007(5)	0.422±0.008(4)	0.164 ± 0.009(9)
birds	0.346±0.026(4)	0.217±0.017(8)	NaN(9)	0.425±0.042(2)	0.412±0.013(3)	**0.437±0.024**(1)	0.294±0.022(5.5)	0.254±0.015(7)	0.294±0.021(5.5)
emotions	**0.805±0.020**(1)	0.695±0.011(6)	0.730±0.010(4)	0.702±0.032(5)	0.668±0.027(7)	0.579±0.007(9)	0.767±0.012(2)	0.732±0.018(3)	0.626±0.037(8)
enron	**0.701±0.006**(1)	0.659±0.007(3)	0.695±0.014(2)	0.558±0.015(5)	0.450±0.011(6)	0.642±0.013(4)	0.063±0.002(8)	0.059±0.004(9)	0.175 ± 0.004(7)
genbase	**0.996±0.003**(2)	0.989±0.006(7.5)	0.994±0.002(5)	**0.996±0.002**(2)	0.989±0.006(7.5)	**0.996±0.003**(2)	0.994±0.005(5)	0.994±0.004(5)	0.892 ± 0.045(9)
languagelog	0.359±0.018(3)	0.304±0.009(5)	0.339±0.012(4)	0.267±0.014(6.5)	0.267±0.011(6.5)	**0.387±0.019**(1)	0.245±0.010(8)	0.371±0.020(2)	0.104 ± 0.006(9)
medical	**0.911±0.017**(1)	0.816±0.009(7)	0.870±0.012(5)	0.834±0.018(6)	0.799±0.024(8)	0.873±0.012(4)	0.887±0.151(2)	0.882±0.015(3)	0.411 ± 0.048(9)
scene	0.864±0.014(2)	0.855±0.011(3)	**0.886±0.013**(1)	0.821±0.016(7)	0.756±0.005(8)	0.840±0.006(6)	0.845±0.008(4)	0.844±0.011(5)	0.715 ± 0.027(9)
Avg rank	1.8750(1)	5.6875(7)	4.0000 (2)	4.5625(4)	6.7500(8)	4.2500(3)	4.9375(6)	4.7500(5)	8.1875(9)
	Micro F1 (↑)								
bibtex	0.285±0.005(5)	0.224±0.007(7.5)	0.370±0.013(2)	**0.487±0.006**(1)	0.356±0.003(3)	0.239±0.010(6)	0.328±0.002(4)	0.181±0.007(9)	0.224±0.021(7.5)
birds	0.119±0.034(4)	0.022±0.022(7)	NaN(8.5)	0.068±0.023(5)	0.344±0.020(2)	0.033±0.011(6)	NA(8.5)	0.124±0.019(3)	**0.413±0.016**(1)
emotions	0.144±0.033(9)	0.454±0.010(6)	0.480±0.012(4)	0.456±0.037(5)	**0.644±0.022**(1)	0.297±0.045(8)	0.605±0.012(3)	0.413±0.030(7)	0.626±0.051(2)
enron	0.546±0.013(2)	0.465±0.018(6)	0.241±0.010(8)	0.479±0.016(5)	0.501±0.010(3)	0.445±0.021(7)	0.483±0.011(4)	0.200±0.012(9)	**0.552±0.022**(1)
genbase	0.988±0.003(2)	0.951±0.009(9)	0.970±0.014(7)	**0.991±0.005**(1)	0.987±0.007(3)	0.980±0.018(4)	0.956±0.036(8)	0.975±0.013(5)	0.972±0.009(6)
languagelog	**0.385±0.014**(1)	0.037±0.014(8)	0.172±0.039(5)	0.205±0.023(4)	0.221±0.009(3)	0.058±0.010 (7)	0.074±0.019(6)	0.028±0.005(9)	0.263±0.008(2)
medical	**0.778±0.025**(1)	0.648±0.026(8)	0.753±0.030(4)	0.755±0.029(3)	0.759±0.026(2)	0.714±0.024(5)	0.708±0.077(6)	0.458±0.034(9)	0.698±0.013(7)
scene	0.541±0.010(8)	0.732±0.014(2)	**0.760±0.008**(1)	0.591±0.026(6)	0.674±0.008(3)	0.588±0.009(7)	0.602±0.015(5)	0.427±0.037(9)	0.617±0.019(4)
Avg rank	4.0000(4)	6.6875(8)	4.9375(5)	3.7500(2)	2.5000(1)	6.2500(7)	5.5625(6)	7.5000(9)	3.8125(3)

Table 7 Friedman statistics F_F on five metrics and the referred critical value at significance level $\alpha = 0.05$. The Friedman statistics for each metric (i.e., F_F) are estimated from the differences of average ranking mentioned from Table 6, whereas the critical value is determined by the count of datasets (i.e., N) and algorithms (i.e., k)

Metrics	F_F	Critical value
Hamming loss	19.9417	2.1087
Ranking loss	31.1417	
One error	22.6583	
Coverage	45.2000	
Average precision	26.9667	
Micro F1	22.6167	

We employ Friedman test [62] to examine whether statistical differences hold for selected evaluation metrics, given the experimental results generated by multiple algorithms across selected datasets. Parameters K and N mean the count of comparing algorithms and datasets, respectively. $R_j = (1/N)\sum_{i=1}^{N} r_i^j$ ranks the average value of the j-th algorithm on all data sets. The null hypothesis (H_0) thinks that all algorithms' performances have no difference statistically. The Friedman statistic F_F obeys the F-distribution that the numerator is $K - 1$ and the denominator is $(K - 1)(N - 1)$:

$$F_F = \frac{(N-1)\chi_F^2}{N(K-1) - \chi_F^2}. \tag{24}$$

where

$$\chi_F^2 = \frac{12N}{K(K+1)}\left[\sum_j R_j^2 - \frac{K(K+1)^2}{4}\right]. \tag{25}$$

Table 7 enumerates the Friedman statistics[62] for all evaluation metrics F_F and the corresponding critical value given the 72 times of five-fold cross-validation (9 comparing algorithms × 8 datasets). On all metrics, the average performance is different statistically. It signifies that the classification performance on all metrics has statistical differences. Therefore, we can further examine the superiority of MGT-LEML over the considered algorithms with some post hoc tests.

We employ the Holm procedure [62] to conduct pairwise comparisons between MGT-LEML (denoted as R_1) and the remaining algorithms (denoted as R_j, where $j = 2, 3, \ldots, 9$). The notation z_j (with $j = 2, 3, \ldots, 9$) records the average ranking of a particular evaluation metric in ascending order and is computed as follows:

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

We use p_j to represent the p-value of z_j. MGT-LEML gains statistical superiority at confidence level $\alpha = 0.05$ if the z_j is smaller than the corresponding p_j. For readability, we highlight those algorithms with bold size from Tables 8, 9, 10, 11, 12.

As shown from Tables 8, 9, 10, 11, 12, and 13, MGT-LEML is statistically superior to algorithm Glocal on metrics *Hamming Loss* and *Coverage*, and is statistically superior to algorithms MLTSVM and MCGM on all metrics except for metrics *One Error* and *Micro F1*, and is statistically superior to algorithm fRA*k*EL on metrics *Ranking Loss* and *Coverage*, and is statistically superior to algorithm LLSF on metrics *Hamming Loss* and *Ranking Loss*, and is statistically superior to algorithms LIFT and HNOML on metrics *Coverage*, and is statistically superior to algorithm ML*k*NN on metrics *Hamming Loss* and *Average Precision*.

5 Discussions

Although the MGT-LEML achieves satisfying classification performance as a whole, the compromised results on *Micro F1* imply that the discrimination on the minor class (i.e., instances with certain classifications) requires improvement. The smaller ranking of MGT-LEML against LLSF on all metrics except *Micro F1* demonstrates that

Table 8 Comparisons of MGT-LEML with remaining algorithms examined by Holm procedure on metric *Hamming Loss*. The algorithms in bold size are statistically inferior to MGT-LEML at significance level $\alpha = 0.05$

j	Algorithm	z_j	p	Holm
2	**MCGM**	−3.742771	0.000182	0.00625
3	**MLTSVM**	−3.377622	0.000731	0.00714
4	**Glocal**	−3.103761	0.001911	0.00833
5	**LLSF**	−2.647326	0.008113	0.01000
6	**ML *k*NN**	−2.556039	0.010587	0.01250
7	HNOML	−2.282177	0.022479	0.01667
8	fRA*kEL*	−2.145247	0.031933	0.02500
9	LIFT	−1.506237	0.132006	0.05000

Table 9 Comparisons of MGT-LEML with remaining algorithms examined by Holm procedure on metric *Ranking Loss*. The algorithms in bold size are statistically inferior to MGT-LEML at significance level $\alpha = 0.05$

j	Algorithm	z_j	p	Holm
2	**MLTSVM**	−4.381780	0.000012	0.00625
3	**fRA *k*EL**	−4.107919	0.000040	0.00714
4	**MCGM**	−3.377622	0.000731	0.00833
5	**LLSF**	−2.875543	0.004033	0.01000
6	Glocal	−2.282177	0.022479	0.01250
7	HNOML	−2.190890	0.028460	0.01667
8	ML*k*NN	−1.688811	0.091256	0.02500
9	LIFT	−1.278019	0.201243	0.05000

Table 10 Comparisons of MGT-LEML with remaining algorithms examined by Holm procedure on metric *One Error*. The algorithms in bold size are statistically inferior to MGT-LEML at significance level $\alpha = 0.05$

j	Algorithm	z_j	p	Holm
2	Glocal	−2.373465	0.017600	0.00625
3	ML*k*NN	−2.327822	0.019900	0.00714
4	MCGM	−2.327822	0.019900	0.00833
5	HNOML	−2.053960	0.040000	0.01000
6	LLSF	−1.095445	0.273300	0.01250
7	LIFT	−0.639010	0.522800	0.01667
8	MLTSVM	−0.274043	0.784100	0.02500
9	fRA*k*EL	0.822130	1.000000	0.05000

Table 11 Comparisons of MGT-LEML with remaining algorithms examined by Holm procedure on metric *Coverage*. The algorithms in bold size are statistically inferior to MGT-LEML at significance level $\alpha = 0.05$

j	Algorithm	z_j	p	Holm
2	**MCGM**	−5.842374	0.000000	0.00625
3	**MLTSVM**	−4.290493	0.000018	0.00714
4	**Glocal**	−3.651484	0.000261	0.00833
5	**LIFT**	−3.468910	0.000523	0.01000
6	**fRA *k*EL**	−3.012474	0.002591	0.01250
7	**HNOML**	−2.464752	0.013711	0.01667
8	ML*k*NN	−1.825742	0.067889	0.02500
9	LLSF	−1.734455	0.082837	0.05000

Table 12 Comparisons of MGT-LEML with remaining algorithms examined by Holm procedure on metric *Average Precision*. The algorithms in bold size are statistically inferior to MGT-LEML at significance level $\alpha = 0.05$

j	Algorithm	z_j	p	Holm
2	**MCGM**	−4.609998	0.000004	0.00625
3	**MLTSVM**	−3.560197	0.000371	0.00714
4	**ML *k*NN**	−2.784256	0.005365	0.00833
5	HNOML	−2.236534	0.025317	0.01000
6	fRA*k*EL	−2.099603	0.035764	0.01250
7	LLSF	−1.962672	0.049684	0.01667
8	Glocal	−1.734455	0.082837	0.02500
9	LIFT	−1.551881	0.120691	0.05000

Table 13 Comparisons of MGT-LEML with remaining algorithms examined by Holm procedure on metric *MicroF1*. The algorithms in bold size are statistically inferior to MGT-LEML at significance level $\alpha = 0.05$

j	Algorithm	z_j	p	Holm
2	fRA*k*EL	−2.556039	0.010600	0.00625
3	ML*k*NN	−1.962673	0.049700	0.00714
4	Glocal	−1.643168	0.100300	0.00833
5	HNOML	−1.141089	0.253800	0.01000
6	LIFT	−0.684653	0.493600	0.01250
7	MCGM	0.136931	1.000000	0.01667
8	LLSF	0.182574	1.000000	0.02500
9	MLTSVM	1.095445	1.000000	0.05000

three-way decisions significantly improve the upper bound of classification accuracy, and such dominance applies to varying domains. However, the effectiveness of employing three-way decisions requires more comprehensive comparisons. Firstly, the settings of six loss functions are from experts, which means the results are only effective and may be sub-optimal. It is worth examining how much the classification performance fluctuates as the loss functions change. Although the latent label correlation is unknown, the relationship between data characteristics and loss functions may reveal some insightful ideas. Secondly, how to leverage the distribution of uncertain classifications to determine the uncertain instance remains an open issue. In our approach, we formulate a classical Top-K problem. Thus, we raise three questions here. (1) How to objectively measure the influence of label correlation; (2) how to determine the optimal components of uncertain instances; and (3) how much performance difference lies between the optimal solution and the presented three-way-based label enhancement schema. We believe there are some trade-off factors between computational efficiency and performance improvement. Thirdly, it is conducive to learning the label-dependent weights from instances of both global and local label correlations. We believe such examinations can facilitate the understanding of multi-label classification. Nevertheless, the MGT-LEML is a promising solution and reveals that leveraging uncertainty is conducive to boosting classification performance.

6 Conclusions

This paper presents a novel label enhancement-based multi-label classification model with multi-granular thresholds. Following the theory of three-way decisions, this model identifies uncertain instances and improves the classification by label enhancement. Results on benchmarks have demonstrated that, with fine-granularity supervision, the reduction of label ambiguity leads to significant improvements in classification performance. For the forthcoming, we will focus on components optimization for both trisecting and acting. In addition, we will examine the performance of MGT-LEML in the specific application domain.

Acknowledgements We gratitude the following fund programs: China Postdoctoral Science Foundation (Grant No. 2022M713491), National Natural Science Foundation of China (Grant No. 61976158, 61976160, 62076182, 62163016, 62006172, 61906137) and Jiangxi "Double Thousand Plan" (Grant No. 20212ACB202001). These funding projects helped in the completion of the article. The authors have no relevant financial or non-financial interests to disclose.

Data availability The datasets generated and analyzed during the current study are available from the corresponding author on reasonable request.

References

1. Zhang ML, Zhou ZH (2014) A review on multi-label learning algorithms. IEEE Trans Knowl Data Eng 26(8):1819–1837
2. Gibaja E, Ventura S (2015) A tutorial on multilabel learning. ACM Comput Surv 47(3):1–38
3. Liu WW, Shen XB, Wang HB, Tsang IW (2022) The emerging trends of multi-label learning. IEEE Trans Pattern Anal Mach Intell 44(3):7955–7974
4. Van Landeghem J, Blaschko M, Anckaert B, Moens MF (2022) Benchmarking scalable predictive uncertainty in text classification. IEEE Access 10:43703–43737
5. Luo LY, Yu LQ, Chen H, Liu QD, Wang X, Xu JQ, Heng PA (2020) Deep mining external imperfect data for chest x-ray disease screening. IEEE Trans Med Imaging 39(11):3583–3594
6. Zhang H, Jiang LX, Xu WQ (2021) Multiple noisy label distribution propagation for crowdsourcing. Paper presented at the proceedings of 28th international joint conference on artificial intelligence, Macao, Peoples R China, 10–16 August 2019
7. Luo JQ, He B, Ou Y (2021) Topic-based label distribution learning to exploit label ambiguity for scene classification. Neural Comput Appl 33(23):16181–16196
8. Li PP, Hu YB, Wu X, He R, Sun ZN (2020) Deep label refinement for age estimation. Pattern Recognit. 100:107178
9. He ZZ, Li X, Zhang ZF, Wu F, Geng X, Zhang YQ, Yang MH, Zhuang YT (2017) Data-dependent label distribution learning for age estimation. IEEE Trans Image Process 26(8):3846–3858
10. Boutell MR, Luo J, Shen X, Brown CM (2004) Learning multi-label scene classification. Pattern Recognit. 37(9):1757–1771
11. Tsoumakas G, Vlahavas I (2007) Random k-labelsets: an ensemble method for multilabel classification. Paper presented at the 18th European conference on machine learning (ECML 2007)/11th European conference on principles and practice of knowledge discovery in databases (PKDD 2007), Warsaw University, Poland, 17–21 September 2007
12. Huang J, Li GR, Huang QM, Wu XD (2016) Learning label-specific features and class-dependent labels for multi-label classification. IEEE Trans Knowl Data Eng 28(12):3309–3323
13. Zhang YJ, Miao DQ, Zhang ZF, Xu JF, Luo S (2018) A three-way selective ensemble model for multi-label classification. Int J Approx Reason 103:394–413
14. Zhang ML, Zhou ZH (2007) Ml-knn: A lazy learning approach to multi-label learning. Pattern Recognit. 40(7):2038–2048
15. Wu QY, Tan MK, Song HJ, Chen J, Ng MK (2016) Ml-forest: a multi-label tree ensemble method for multi-label classification. IEEE Trans Knowl Data Eng 28(10):2665–2680
16. Chen YH, Shao C, Li N, Deng NY (2016) Mltsvm: a novel twin support vector machine to multi-label learning. Pattern Recognit. 52:61–74
17. Xu SP, Yang XB, Yu HL, Yu DJ, Yang JY, Tsang ECC (2016) Multi-label learning with label-specific feature reduction. Knowl-Based Syst 104:52–61
18. Geng X (2016) Label distribution learning. IEEE Trans Knowl Data Eng 28(7):1734–1748
19. Xu SP, Ju HR, Shang L, Pedrycz W, Yang XB, Li C (2020) Label distribution learning: a local collaborative mechanism. Int J Approx Reason 121:59–84
20. Xu SP, Shang L, Shen FR (2017) Latent semantics encoding for label distribution learning. Paper presented at the 28th international joint conference on artificial intelligence (IJCAI'19), China, Macao, Aug 2019. p 3982–3988
21. Tao A, Xu N, Geng X (2018) Labeling information enhancement for multi-label learning with low-rank subspace. Paper presented at the 15th Pacific Rim international conference on artificial intelligence (PRICAI), Nanjing, Peoples R China, 28–31 Aug 2018
22. Li YK, Zhang ML, Geng X (2015) Leveraging implicit relative labeling importance information for effective multi-label learning. Paper presented at the IEEE international conference on data mining (ICDM), Atlantic City, NJ, 14–17 Nov 2015
23. Xu N, Tao A, Geng X (2018) Label enhancement for label distribution. Paper presented at the 27th international joint conference on artificial intelligence (IJCAI), Stockholm, Sweden, 13–19 Jul 2018
24. Shao RF, Xu N, Geng X (2018) Multi-label learning with label enhancement. Paper presented at the 18th IEEE international conference on data mining workshops (ICDMW), Singapore, Singapore, 17–20 Nov 2018
25. Xu WH, Guo DD, Qian YH, Ding WP (2022) Two-way concept-cognitive learning method: A fuzzy-based progressive learning. IEEE Trans. Fuzzy Syst. 1–15. https://doi.org/10.1109/TFUZZ.2022.3216110
26. Xu WH, Yuan KH, Ding WP (2023) An emerging fuzzy feature selection method using composite entropy-based uncertainty measure and data distribution. IEEE Trans. Emerg. Top Comput. Intell. 7(1):76–88
27. Xu WH, Pan YZ, Chen XW, Ding WP, Qian YH (2022) A novel dynamic fusion approach using information entropy for interval-valued ordered datasets. IEEE Trans. Big Data. https://doi.org/10.1109/TBDATA.2022.3215494
28. Yao YY (2009) Three-way decision: an interpretation of rules in rough set theory. Paper presented at the 4th international conference on rough sets and knowledge technology (RSKT), Gold Coast, Australia, 14–16 Jul 2009
29. Yao YY (2018) Three-way decision and granular computing. Int J Approx Reason 103:107–123
30. Yao YY (2020) Tri-level thinking: models of three-way decision. Int J Mach Learn Cybern 11:947–959
31. Yao YY (2021) The geometry of three-way decision. Appl Intell 51(9):6298–6325

32. Zhang XY, Gou HY, Lv ZY, Miao DQ (2021) Double-quantitative distance measurement and classification learning based on the tri-level granular structure of neighborhood system. Knowl-Based Syst 217:106799
33. Zhang K, Dai JH, Zhan JM (2021) A new classification and ranking decision method based on three-way decision theory and topsis models. Inf Sci 568:54–85
34. Liu JB, Li HX, Huang B, Liu Y, Liu D (2021) Convex combination-based consensus analysis for intuitionistic fuzzy three-way group decision. Inf Sci 574:542–566
35. Liang DC, Fu YY, Xu ZS (2022) Three-way group consensus decision based on hierarchical social network consisting of decision makers and participants. Inf Sci 585:289–312
36. Xu WH, Guo DD, Mi JS, Qian YH, Zheng KY, Ding WP (2023) Two-way concept-cognitive learning via concept movement viewpoint. IEEE Trans. Neural Netw. Learn. Syst. 1–15. https://doi.org/10.1109/TNNLS.2023.3235800
37. Yuan KH, Xu WH, Li WT, Ding WP (2022) An incremental learning mechanism for object classification based on progressive fuzzy three-way concept. Inf Sci 584:127–147
38. Ren FJ, Wang L (2017) Sentiment analysis of text based on three-way decisions. J. Intell. Fuzzy. Syst. 33(1):245–254
39. Zhang YJ, Zhao TN, Miao DQ, Pedrycz W (2022) Granular multilabel batch active learning with pairwise label correlation. IEEE. Trans. Syst. Man. Cybern. -Syst. 52(5):3079–3091
40. Qian WB, Huang JT, Wang YL, Xie YH (2021) Label distribution feature selection for multi-label classification with rough set. Int J Approx Reason 128:32–55
41. Zhang ML, Wu L (2015) Lift: Multi-label learning with label-specific features. IEEE Trans Pattern Anal Mach Intell 37(1):107–120
42. Zhu Y, Kwok JT, Zhou ZH (2018) Multi-label learning with global and local label correlation. IEEE Trans Knowl Data Eng 30(6):1081–1094
43. Zhang J, Luo ZM, Li CD, Zhou CG, Li SZ (2019) Manifold regularized discriminative feature selection for multi-label learning. Pattern Recognit. 95:136–150
44. Min XY, Qian K, Zhang BW, Song GJ, Min F (2022) Multi-label active learning through serial-parallel neural networks. Knowl-Based Syst 251:109226
45. Kimura K, Kudo M, Sun L, Koujaku S (2017) Fast random k-labelsets for large-scale multi-label classification. Paper presented at the 23rd international conference on pattern recognition (ICPR), Mexican Assoc Comp Vis Robot & Neural Comp, Mexico, 04–08 Dec 2016
46. Nazmi S, Yan XY, Homaifar A, Docuettee E (2020) Evolving multi-label classification rules by exploiting high-order label correlations. Neurocomput. 417:176–186
47. Wang, R., Kwong, S., Wang, X., Jia, Y.: Active k-labelsets ensemble for multi-label classification **109**, 107583 (2021)
48. Huang J, Qin F, Zheng X, Cheng ZK, Yuan ZX, Zhang WG, Huang QM (2019) Improving multi-label classification with missing labels by learning label-specific features. Inf Sci 492:124–146
49. Lin YJ, Hu QH, Liu JH, Zhu XQ, Wu XD (2021) Mulfe: multi-label learning via label-specific feature space ensemble. ACM Trans Knowl Discov Data 16(1):5
50. Cheng YS, Zhang C, Pang SF (2022) Multi-label space reshape for semantic-rich label-specific features learning. Int J Mach Learn Cybern 13:1005–1019
51. Wang YB, Hang JY, Zhang ML (2022) Stable label-specific features generation for multi-label learning via mixture-based clustering ensemble. IEEE-CAA J. Automatica Sin. 9(7):1248–1261
52. Long XD, Qian WB, Wang YL, Shu WH (2021) Cost-sensitive feature selection on multi-label data via neighborhood granularity and label enhancement. Appl Intell 51(4):2210–2232
53. Liu XY, Zhu JH, Zheng QH, Li ZY, Liu RX, Wang J (2021) Bidirectional loss function for label enhancement and distribution learning. Knowl-Based Syst 213:106690
54. Liu XY, Zhu JH, Li ZY, Tian ZQ, Jia XY, Chen L (2021) Unified framework for learning with label distribution. Inf. Fusion. 75:116–130
55. Li WW, Chen J, Gao PX, Huang ZQ (2022) Label enhancement with label-specific feature learning. Int J Mach Learn Cybern 13(10):2857–2867
56. Qian WB, Xiong CZ, Qian YH, Wang YL (2022) Label enhancement-based feature selection via fuzzy neighborhood discrimination index. Knowl-Based Syst 250:109119
57. Tan C, Chen S, Ji GL, Geng X (2022) Multilabel distribution learning based on multioutput regression and manifold learning. IEEE Trans. Cybern. 52(6):5064–5078
58. Tsoumakas G, Spyromitros-Xiousfis E, Vilcke I (2011) Mulan: a java library for multi-label learning. J Mach Learn Res 12(7):2411–2414
59. Zhang CQ, Yu ZW, Fu HZ, Zhu PF, Chen L, Hu QH (2020) Hybrid noise-oriented multilabel learning. IEEE Trans. Cybern. 50(6):2837–2850
60. Ma JH, Chiu BCY, Chow TWS (2022) Multilabel classification with group-based mapping: a framework with local feature selection and local label correlation. IEEE Trans. Cybern. 52(6):4596–4610
61. Schapire R, Singer Y (2000) A boosting-based system for text categorization. Mach Learn 39(2/3):135–168
62. Demsar J (2006) Statistical comparisons of classifier over multiple data sets. J Mach Learn Res 7:1–30

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.