International Journal of Approximate Reasoning 150 (2022) 172–187

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

Selective label enhancement for multi-label classification based on three-way decisions

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

[a] *Department of Computer Science and Technology, Tongji University, Shanghai, 201804, China*
[b] *Key Laboratory of Embedded System and Service Computing, Ministry of Education, Tongji University, Shanghai, 201804, China*
[c] *China UnionPay Co., Ltd, Shanghai, 201201, China*
[d] *Postdoctoral Research Station of Computer Science and Technology, Fudan University, Shanghai, 200433, China*
[e] *Department of Electrical and Computer Engineering, University of Alberta, Edmonton, T6R 2V4, Canada*
[f] *System Research Institute, Polish Academy of Sciences, Warsaw, PL-01447, Poland*

ARTICLE INFO

Article history:
Received 6 February 2022
Received in revised form 23 July 2022
Accepted 11 August 2022
Available online 23 August 2022

Keywords:
Multi-label classification
Three-way decisions
Uncertainty
Trisecting-acting-outcome model
Label enhancement

ABSTRACT

Multi-label classification is a challenging issue in the data science community due to the ambiguity of label semantics. Existing studies mainly focus on improving label association with logical labels, but the performance suffers from the threshold setting. Although label distribution learning gains superior discrimination, the expenditure of collecting large-scale fine-grained numerical labels is intolerable. To address the uncertainty of logical label semantics, we propose a novel model called three-way decisions with label enhancement (3WDLE). For unseen instances, we implement a trisecting-acting-outcome framework. In the trisecting stage, an uncertainty measure called global uncertain-prone degree partitions these instances into uncertain and certain regions, where the trisecting procedure is completed from label level to instance level by leveraging the distributions of pseudo-label information. In the acting stage, instances recognized as certain regions directly take the results generated by label-specific learning, whereas the remaining are reclassified by conducting selective label enhancement. The enriched knowledge generated by the label enhancement module is learnt on trustworthy instances only. In the outcome stage, we adopt five evaluation metrics to evaluate the classification performance from the perspectives of both labels and instances. In this way, three-way decisions provide a systematic methodology to deal with uncertainty in multi-label classification, which combines logical label learning with numerical label learning into a unified framework to optimize the performance of the multi-label classification model. Extensive experiments demonstrate the superiority of 3WDLE over state-of-the-art multi-label classifications with logical labels only.

1. Introduction

Data science is a discipline which emphasizes learning essential knowledge from observed data with scientific methodology, and classification is one of the influential research directions. It is a procedure aiming at classifying unseen instances

[*] Corresponding author at: China UnionPay Co., Ltd, Shanghai, 201201, China.
[**] Corresponding author at: Department of Computer Science and Technology, Tongji University, Shanghai, 201804, China.
E-mail addresses: zhangyuanjian@unionpay.com (Y. Zhang), dqmiao@tongji.edu.cn (D. Miao).

https://doi.org/10.1016/j.ijar.2022.08.008
0888-613X/© 2022 Elsevier Inc. All rights reserved.

with the learnt knowledge towards the intention among concepts. To computationally complete the classification, people introduce artificially annotated labels as the carrier of concept extension. Traditionally, explore the “is-a” relationship between instances and labels, which means each instance is associated with only one label. Under the single-label context, the relationships among classes are mutually exclusive.

Multi-label annotations [1–4] introduce multiple labels to describe instances simultaneously, a semantical extension of the single-label form. Multi-label classification learns a real-valued mapping from feature space to label space which determines the label associations of unseen instances. In its early stage, the relationship between an instance and a label is logical, i.e., it either holds or does not hold. For example, facial expressions are a mixture of *sad*, *anger*, and *disgust*. Although a given facial expression may express different emotions, the facial expression on a specific emotion annotation (e.g., *sad*) should be either associated or not. Complex applications such as psychological consulting focus on how many degrees an emotion describes the facial expression. Thus, the rationale for such multi-label classification arises from logical labels. Different domains apply this assumption, such as smart grid management [5], disease diagnosis [6], and image classification [7]. Recent years have witnessed a surge of multi-label classification algorithms, and we categorize these as *problem transformation* and *algorithm adaptation* in terms of problem-solving structure. The previous one transforms multi-label classification into a collection of simplified learning scenarios like single-label classification. Representative work involves Binary Relevance [8], Random *k*-labelsets [9], and LLSF [10]. In contrast, the latter extends the existing algorithms to a multi-output form. Representative work involves BP-MLL [11], ML-*k*NN [12], ML-Forest [13], and MLTSVM [14]. These algorithms endeavor to boost the classification performance by learning the latent label correlations, and label-specific learning is a branch that characterizes the semantics difference of labels.

The capability of information processing on multi-label does not merely halt at determining whether the logical relationship between an instance and an arbitrary label holds. The multi-label implies interactions between different classes, spontaneously arising from the requirement for complex applications like situational awareness and psychological consulting. In this case, it is inadequate to solve the problem if we only have qualitative labels. Currently, there are two types of numerical labels. The first category regards numerical labels as low-rank latent representations of observed logical labels [15–18]. While obtaining robustness on low-quality multi-label classification, the label discrimination is not essentially improved and is beyond the scope of this paper. The second category argues that the extension from logical to numerical labels can endow labels with the capability to describe instances [19–22]. With the same label dimensionality as logical labels, the description degrees of all labels constitute a real-valued vector and define the process of fitting the observed vector as label distribution learning [23]. In label distribution learning, a facial expression can be a combination of *slight sadness*, *some anger*, and *a bit of disgust*. Although it offers more discriminative information than logical labels on many domains, it is more costly to assign accurate numerical labels. The bottleneck is two-fold. It not only requires a comprehensive understanding of all labels but also requires a discriminative capability among labels. Sophisticated annotation requires considerable expenditure on both budget and time. An alternative solution is to learn the numerical labels from the logical labels called *label enhancement* [24–27].

From the cognition viewpoint, we resort to more powerful knowledge only if the uncertainty is large enough and make decisions otherwise. We do not complete the decision process for complex problem-solving in one step, and each decision implies a divide-and-conquer procedure driven by an uncertainty measure. The mechanism is unavailable in the existing label distribution learning methods and requires a novel decision-making theory.

Three-way decisions [28,29] (a.k.a. TAO model [30,31]) provide an intuitive interpretation of three regions concerning concepts with roughness via a problem-specific evaluation function. The deduced positive, negative, and boundary regions are an implementation of thinking in threes. Many scholars broadened the context of the decision by employing the three-way decisions on data mining-related topics like in attribute reduction [32–34], concept analysis [35–37] and clustering [38–40]. With superior performance on effectiveness and efficiency, it is an emerging decision theory for problem-solving with uncertainty [41–46]. TAO is an abbreviation of trisecting, acting and outcome. In particular, the trisecting step concentrates on the tripartition of the entire universal driven by uncertainty measures; the acting emphasizes adopting actions on the deduced three parts; the outcome step aims at evaluating the quality of trisecting and acting. The three steps (i.e., trisecting, acting and outcome) constitute an atomic routine for the viewpoint towards uncertainty and continue until the uncertainty is negligible.

Recent years have witnessed some progress on multi-label classification with three-way decisions [47–52]. Three-way decisions achieve impressive performance in multi-label classifications, especially for cases like supervised multi-label learning [47,48,52], incremental multi-label learning [49], active multi-label learning [50], and multi-view weak multi-label learning [51]. However, the referred uncertainty is within either logical labels or numerical labels. They consider either feature integration or algorithm integration to deal with label ambiguity, whereas the semantics integration of labels remains untouched.

This paper presented a three-way label enhancement model (3WDLE) for multi-label classification. We assume that the degenerated performance of multi-label classification stems from uncertain instances, and it is critical to identify these instances. Concretely, the instance trisecting is based on the global uncertain-prone degree measures and implemented via an uncertainty matrix. The uncertainty matrix preserves the qualitative distribution of unreliable classifications across the label space and develops from label level to instance level. Two facets, the relative distance to the label-specific hyperplane and the proportion of heterogeneous instances within the neighborhood, are considered to construct the element in the uncertainty matrix. By analyzing the distribution of unreliable classifications among the uncertainty matrix, we select the instances misclassified among most labels. The enhanced model strengthens classifications of these instances, which constructs the label enhancement on certain-prone training instances by automatically leveraging the feature topology. Compared with the existing multi-label classifications models, our contributions are enumerated as follows:

(1) We unify logical label learning and label enhancement learning under the classification uncertainty to promote multi-label classification for the first time. The global uncertain-prone degree of instance measures the uncertainty and selects instances for further processing, contributing to the label information mining of the label enhancement.

(2) The determination of the label association for unseen instances follows the three-way decisions theory. In trisecting stage, we conduct instance trisecting on unseen instances and determine the uncertainty instances with greater possibilities of misclassification across most labels in a bottom-up manner. In the acting stage, a more discriminative module called label enhancement strengthens the classifications of selected uncertainty instances, whereas taking those pseudo-labels learnt from a logical label-based model for the remaining. The label enhancement is trained on trustworthy instances to strengthen the discrimination on each label. We realize the outcome stage by evaluating five multi-label classification metrics.

(3) The 3WDLE integrates the merits of both label-specific learning and label enhancement. Apart from the instance trisecting, the remaining are all one-stage models. It not only preserves the characteristics of each label but also reduces the sophisticated knowledge of label importance.

The remainder of the paper is structured as follows. Section 2 reviews some preliminaries on label-specific learning and label enhancement. Section 3 elaborates on the formulation of the proposed 3WDLE. Section 4 analyzes the experimental results. Section 5 further discusses some characteristics of 3WDLE. Section 6 concludes this work and identifies future directions.

2. Preliminaries

This section briefly reviews preliminary studies on label-specific feature learning and label enhancement.

2.1. Label-specific feature learning

Label-specific feature [10,53–55] assumes that each label has its unique characteristics and describes it by different feature subsets. For the simplicity of computation, the learning label-specific feature (LLSF) [10] considers second-order label correlation (a.k.a. pairwise, one label is at most dependent on another) and assumes three hypotheses:

(1) Discrimination: The set of i-th label-specific features should be most pertinent to the corresponding label (l_i), and the included components should be different from other label-specific features;

(2) Sparsity: The label-specific features should be sparse as compared with feature space;

(3) Sharability: The number of the same features of two label-specific features with strong label correlations should be more than those with weak label correlations.

Based on the previous hypotheses, formulate the objective function as:

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

where $\mathbf{X}$ and $\mathbf{Y}$ represent the features and labels of multi-label data, $\mathbf{W}=[\mathbf{w}_1,\mathbf{w}_2,\ldots,\mathbf{w}_i,\ldots,\mathbf{w}_m]$, and the basic element $\mathbf{w}_i$ represents the weight of label-specific feature of l_i, which is composed of non-zero terms in $\mathbf{w}_i$. $\|\cdot\|_F^2$ denotes the Frobenius norm. $\mathbf{R}=\left[r_{ij}\right]$ represents a matrix composed of second-order label relevance. $r_{ij}=1-c_{ij}$, where c_{ij} measures the correlation between label l_i and label l_j. The label correlation is computed with cosine similarity. $tr(\cdot)$ denotes the matrix traces. Symbols δ and η are the balance factors. The inner product of $\mathbf{w}_i$ and $\mathbf{w}_j$ describes the correlation between label l_i and label l_j from feature view. The stronger the correlation is, the larger the inner product becomes, and vice versa.

For the prediction of unseen instances, LLSF employs logistic regression and can be denoted as:

$$\hat{\mathbf{Y}}=sgn\left(\mathbf{XW}-\tau\right), \tag{2}$$

where $sgn(\cdot)$ returns 1 if the condition holds, and 0 otherwise.

2.2. Label enhancement

Label enhancement assumes that the labels of each instance are intrinsical with real-valued labels. Therefore, latent labels can be recovered from qualitative logical labels ($\mathbf{Y}$) to the quantitative numerical labels ($\mathbf{U}$) via instance-level or label-level smoothness [20,27,56–59]. The distribution of numerical labels describes the relative importance of different labels in describing a given instance.

Three hypotheses are presented in label enhancement multi-label learning (LEMLL) [27] to guarantee effectiveness.

(1) Linear relevance: The mapping from feature space to numerical label $g:\mathbf{X}\rightarrow\mathbf{U}$ follows a linear model;

(2) Label similarity: The value of the learned numerical label should be close to the value of the original logical label;

(3) Topology consistency: The instances with similar features share similar numerical label values.

Based on the previous assumptions, we formulate the objective function as:

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

where $\sum_{i=1}^{n} L_R\left(R_i\right)$ denotes the loss function from feature space to numerical labels, with the regularizer as $R_i = \left\|\xi_i\right\|_2 = \sqrt{\xi_i^\top \xi_i}$, where $\xi_i = \mathbf{u}_i - \boldsymbol{\Theta}\varphi\left(\mathbf{x}_i\right) - \mathbf{b}$, and $\varphi\left(\mathbf{x}_i\right)$ is a mapping from instance $\mathbf{x}_i$ to a high-dimensional space $\mathbb{R}^{\mathbb{H}}$; $\boldsymbol{\Theta}$ and $\mathbf{b}$ are the parameters in linear regression model. $\left\|\mathbf{U}-\mathbf{Y}\right\|_F^2$ is the implementation of label similarity assumption measured by Frobenius norm (abbreviated as F). α, β, γ are all balance factors. $tr\left(\mathbf{U}^\top \mathbf{M}\mathbf{U}\right) = \left\|\mathbf{U}-\mathbf{W}\mathbf{U}\right\|_F^2$ is the implementation of topology consistency, where $tr\left(\cdot\right)$ represents matrix trace, and $\mathbf{M} = \left(\mathbf{I}-\mathbf{W}\right)^\top \left(\mathbf{I}-\mathbf{W}\right)$, with $\mathbf{I}$ be an identity matrix and $\mathbf{W}$ be the weight matrix constructed by fully connected graph $G = \left(\mathbf{V}, \mathbf{E}, \mathbf{W}\right)$, which describes the associations on arbitrary two instances.

LEMLL leverages a kernel logistic regression and definites the prediction of unseen instances as:

$$\hat{\mathbf{Y}} = sgn\left(\boldsymbol{\Theta}\varphi\left(\mathbf{X}\right) + \mathbf{b} - \tau\right). \tag{4}$$

3. The 3WDLE model

3.1. Basic idea

The 3WDLE model is an implementation of the TAO model for multi-label classification. Fig. 1 illustrates the pipeline of 3WDLE on unseen instances. The entire model is composed of three components, i.e., trisecting, acting, and outcome. For unseen instances $\mathbf{X}_2$, we conduct instance trisecting based on an uncertainty measure called misclassification degree. The uncertain instances $\mathbf{X}_2^{(d,\rho)}$ correspond to the deferment region of three-way decisions strengthened by a selective label enhancement module denoted by $f_2\left(\cdot\right)$. The label-specific learning model denoted by $f_1\left(\cdot\right)$ determines the remaining. The conventional metrics evaluate the performance of uncertain and certain instances in an unseen instance set from both label-based and example-based perspectives. In what follows, we elaborate on the instance trisecting and construction of $f_2\left(\cdot\right)$, which highlights with a yellow box in trisecting stage and acting stage, respectively. Formally speaking, we definite the final classification of $\mathbf{X}_2$ (denoted as $\hat{\mathbf{Y}}_2^*$) as:

$$\hat{\mathbf{Y}}_2^* = \begin{cases} f_2(\mathbf{X}_2^{(d,\rho)}), & \mathbf{x}_j \in \mathbf{X}_2^{(d,\rho)}; \\ f_1(\neg\mathbf{X}_2^{(d,\rho)}), & \mathbf{x}_j \in \neg\mathbf{X}_2^{(d,\rho)}, \end{cases} \tag{5}$$

where $f_1(\neg\mathbf{X}_2^{(d,\rho)})$ and $f_2(\mathbf{X}_2^{(d,\rho)})$ refer to the results of directly classified instances and the results of deferred instances, respectively.

3.2. Trisecting: instance trisecting

The instance trisecting identifies the instances with a larger possibility of misclassification across most labels. Fig. 2 describes the processing of instance trisection. Taking a problem transformation view, we consider two factors to estimate the uncertainty degree for each unseen instance $\mathbf{x}_j$ on an arbitrary label l_c. One factor is the relative distance from instances to the hyperplane (d), and the other is the proportion of heterogeneous instances within the neighborhood (ρ). The combination of two factors yields local uncertain-prone instances, denoted as $\mathbf{X}_2^{(d_c,\rho_c)}, \forall l_c$. We select global uncertainty instances $\mathbf{X}_2^{(d,\rho)}$ by ranking the average uncertainty degree across label space.

3.2.1. Label-level uncertainty matrix construction

The instance trisecting at the label level intends to figure out the instances with relatively large possibilities of misclassification on an arbitrary label. With label-specific features and pseudo-labels generated by LLSF, we examine the following two factors:

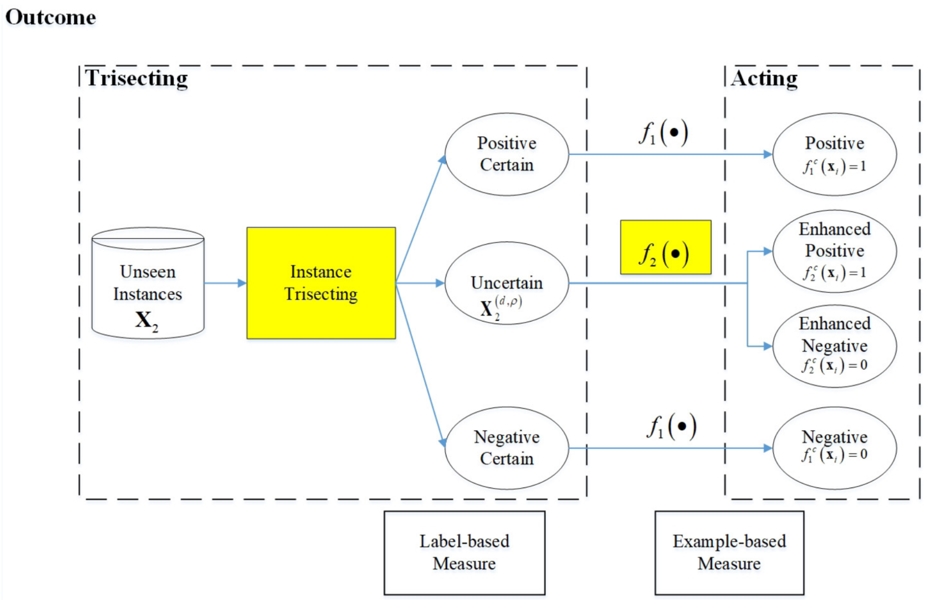
Outcome
Trisecting
Acting
Unseen Instances $\mathbf{X}_2$
Instance Trisecting
Positive Certain
Uncertain $\mathbf{X}_2^{(d,\rho)}$
Negative Certain
$f_1(\bullet)$
$f_2(\bullet)$
$f_1(\bullet)$
Positive $f_1^c(\mathbf{x}_i)=1$
Enhanced Positive $f_2^c(\mathbf{x}_i)=1$
Enhanced Negative $f_2^c(\mathbf{x}_i)=0$
Negative $f_1^c(\mathbf{x}_i)=0$
Label-based Measure
Example-based Measure

Fig. 1. Pipeline of 3WDLE. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

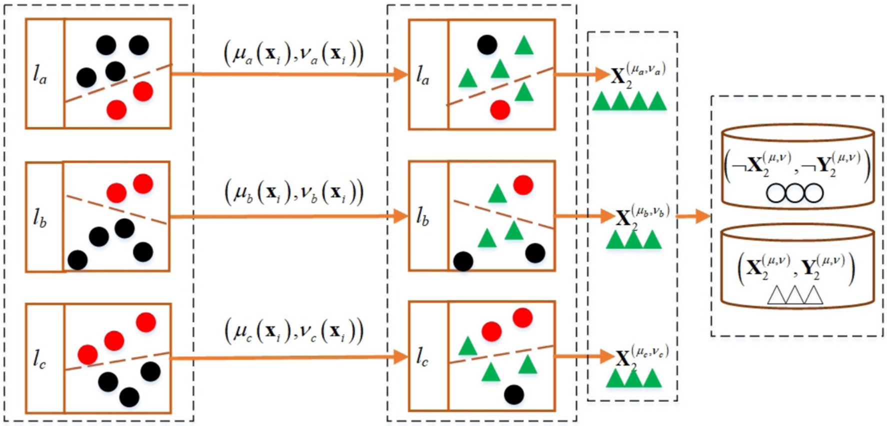
l_a
l_b
l_c
$(\mu_a(\mathbf{x}_i),\nu_a(\mathbf{x}_i))$
$(\mu_b(\mathbf{x}_i),\nu_b(\mathbf{x}_i))$
$(\mu_c(\mathbf{x}_i),\nu_c(\mathbf{x}_i))$
l_a
l_b
l_c
$\mathbf{X}_2^{(\mu_a,\nu_a)}$
$\mathbf{X}_2^{(\mu_b,\nu_b)}$
$\mathbf{X}_2^{(\mu_c,\nu_c)}$
$(\neg\mathbf{X}_2^{(\mu,\nu)},\neg\mathbf{Y}_2^{(\mu,\nu)})$
$(\mathbf{X}_2^{(\mu,\nu)},\mathbf{Y}_2^{(\mu,\nu)})$

Fig. 2. Illustration of instance trisecting: We use six instances and three labels to explain our idea on $\mathbf{X}_2$. After employing $f_1(\cdot)$ on all training instances (i.e., $\mathbf{X}_1$), we employ (d_c,ρ_c) for arbitrary label l_c. The red blocks and black circles refer to the instances that belong to positive and negative with less uncertainty, whereas the green triangles represent local uncertain-prone instances.

(i) relative distance from instance $\mathbf{x}_i$ to hyperplane generated by $f_1(\cdot)$ for each label l_c (denoted as d_{ic});
(ii) the proportion of heterogeneous instances in the kNN of the instance $\mathbf{x}_i$ for each label l_c (denoted as ρ_{ic}).

With label-specific features and pseudo-labels generated by LLSF, we examine the relative distance from instance $\mathbf{x}_i$ to hyperplane generated by $f_1(\cdot)$ for each label l_c (denoted as d_{ic}). The relative distance is defined as:

$$d_{ic} = \left|\hat{y}_{ic} - 0.5\right|, \tag{6}$$

where 0.5 is the default calibrated threshold in LLSF. The smaller the d_{ic} is, the less credible the annotation $\mathbf{x}_i$ on l_c becomes, and vice versa. Hence, the uncertain instances with positive ($\mathbf{X}_2^{c+}$) and negative ($\mathbf{X}_2^{c-}$) pseudo-labels are denoted as:

$$\mathbf{X}_2^{c+} = \left\{\mathbf{x}_i \left|\mathbf{x}_i \in \mathbf{X}_2 \wedge \hat{y}_{ic} = 1 \wedge d_{ic} \leqslant r_c^+\right.\right\}, \tag{7}$$

$$\mathbf{X}_2^{c-} = \left\{\mathbf{x}_i \left|\mathbf{x}_i \in \mathbf{X}_2 \wedge \hat{y}_{ic} = 0 \wedge d_{ic} \geqslant r_c^-\right.\right\}, \tag{8}$$

where r_c^+ and r_c^- denote the average relative distance of testing instances with positive pseudo-label class and negative pseudo-label class on label l_c to the hyperplane, respectively.

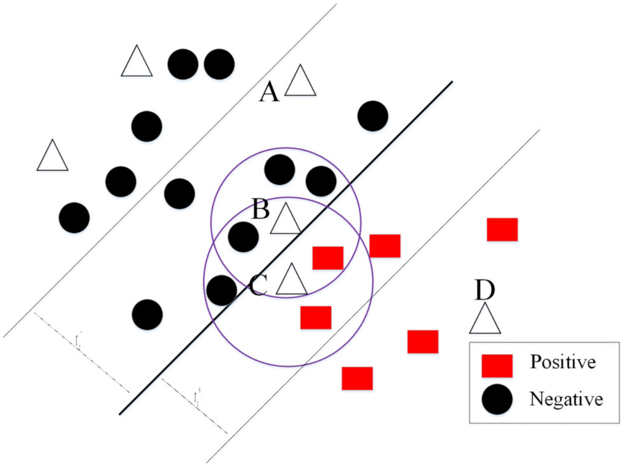

Fig. 3. Illustration for the rationality of the local uncertainty degree: Let the black circles and red blocks denote the instances in $\mathbf{X}_1$ with pseudo-negative and pseudo-positive labels on label l_c, respectively. The solid line represents the classification hyperplane, whereas the two dashed lines represent the plane formed by the average margin (represented as r_c^- and r_c^+) of the negative and positive instances. The six triangles represent six unseen instances in $\mathbf{X}_2$. For readability, we show the neighbors of two instances (i.e., B and C), one can conclude that, given $k=5$, $u_{Bc}>0$ and $u_{Cc}>0$, whereas $u_{Ac}=0$ and $u_{Dc}=0$.

We estimate the misclassification possibility of boundary instances at the label level by calculating the proportion of heterogeneous instances within kNN. The optimum hyperplane learns the semantics of the label, and instances with homogeneous classes are close to each other. Given a hyperplane, an instance is prone to be misclassified if its neighborhood contains many instances with heterogeneous pseudo-labels and is correctly classified otherwise. We compute the proportion of heterogeneous instances within the kNN of $\mathbf{x}_i$ on label l_c (denoted as ρ_{ic}) as follows:

$$\rho_{ic}=\begin{cases}\frac{h_{ic}}{k}\times prior(l_c), & \hat{y}_{ic}=1\wedge\mathbf{x}_i\in\mathbf{X}_2^{c+};\\ \frac{h_{ic}}{k}\times(1-prior(l_c)), & \hat{y}_{ic}=0\wedge\mathbf{x}_i\in\mathbf{X}_2^{c-},\end{cases}\tag{9}$$

where h_{ic} represents the count of instances with heterogeneous class in training set (i.e., $\mathbf{X}_1$) against the instance $\mathbf{x}_i$ on label l_c within k nearest neighbor, and k denotes the neighborhood size. The weight $prior(l_c)=\frac{|y_{ic}=1|+1}{|\mathbf{X}_1|+1}$ is the smoothed prior probability of instances with label l_c in training set.

By integrating d_{ic} with ρ_{ic}, we define a novel metric called local uncertainty degree, u_{ic}, which is computed as:

$$u_{ic}=\frac{\rho_{ic}}{d_{ic}},\tag{10}$$

where $j=1,2,\ldots,s$ and $c=1,2,\ldots,m$. Fig. 3 clearly explains the rationality of proposed uncertainty degree.

By concatenating all label-level uncertainty degrees, the uncertainty matrix U is formally formulated as:

$$U=\begin{bmatrix}u_{11} & u_{12} & \cdots & u_{1m}\\ u_{21} & u_{22} & \cdots & u_{2m}\\ \vdots & \vdots & \ddots & \vdots\\ u_{s1} & u_{s2} & \cdots & u_{sm}\end{bmatrix}_{s\times m},\tag{11}$$

where $s=\sum_c\left|\mathbf{X}_2^{c+}\cup\mathbf{X}_2^{c-}\right|$ denotes the candidates count of instances doing label enhancement. In what follows, we stipulate the uncertainty matrix for the testing set as U, and define $\mathbf{X}_2^{(d_c,\rho_c)}$ according to the instances with non-zero elements in U, i.e.:

$$\mathbf{X}_2^{(d_c,\rho_c)}=\left\{\mathbf{x}_j\left|\mathbf{x}_j\in\mathbf{X}_2^{c+}\cup\mathbf{X}_2^{c-}\wedge u_{jc}>0\right.\right\}.\tag{12}$$

3.2.2. Instance level: uncertainty instance generation

We have obtained an uncertainty matrix with non-zero elements as instances with local uncertain-prone and zero elements as certainty flags. The rows of U reveal the uncertainty distribution of all instances across labels, whereas the columns of U measure the local uncertain-prone degree on instances. Recall that we need to determine the components of global uncertainty instances, and we raise two questions:

(i) How to compute the uncertainty degrees of different instances across all labels?
(ii) How many instances are required to be further processed by label enhancement?

For the first question, we analyze from the perspective of uncertainty matrix rows. The global uncertain-prone degree of instance $\mathbf{x}_j \in \mathbf{X}_2$ is defined as:

$$u_j = \sum_{g=1}^{m} u_{jg}, \tag{13}$$

where $j = 1, 2, \ldots, s$. The larger the u_j is, the larger the overall uncertainty of $\mathbf{x}_j$ becomes. The classification accuracy will be significantly improved if further processing like label enhancement works on such $\mathbf{x}_j$. Therefore, we obtain the global uncertainty ranking of $\mathbf{X}_2$ (denoted as $\acute{\mathbf{X}}_2$) as:

$$\acute{\mathbf{X}}_2 = \left(\mathbf{x}_{\acute{j}_1}, \mathbf{x}_{\acute{j}_2}, \ldots, \mathbf{x}_{\acute{j}_{n_2}}\right), \tag{14}$$

where $u_{\acute{j}_t} \geqslant u_{\acute{j}_{t+1}}$ and $t = 1, 2, \ldots, n_2 - 1$. The higher the order of $\mathbf{x}_j$ in $\acute{\mathbf{X}}_2$ (denoted as $rank(\mathbf{x}_j)$) is, the more likely the $\mathbf{x}_j$ conducts label enhancement.

We calculate the average uncertainty count for the testing set (denoted as n_2^{le}) to the number of instances for label enhancement. It averages the number of local uncertain-prone instances in all labels with local uncertain-prone instances and for the testing set, denoted as:

$$n_2^{le} = \left\lfloor \frac{\sum_{g=1}^{m} \sum_{j=1}^{n_2} sgn\left(u_{jg}\right)}{\sum_{g=1}^{m} sgn\left(\sum_{j=1}^{n_2} u_{jg}\right)} \right\rfloor, \tag{15}$$

where n_2 denotes the actual instances count that participated in the testing by LLSF, and m denotes the label count. It is straightforward to conclude that $0 \leqslant n_2^{le} \leqslant s \leqslant n_2$. Specifically, n_2^{le} reaches minimum if $\forall u_{jg} = 0$ holds, whereas n_2^{le} reaches maximum if $\forall u_{jg} > 0$ holds.

With global uncertainty ranking of $\mathbf{X}_2$ (i.e., $\acute{\mathbf{X}}_2$), and the corresponding average uncertainty count (i.e., n_2^{le}), we select the top n_2^{le} from $\acute{\mathbf{X}}_2$ as $\mathbf{X}_2^{(d,\rho)}$, denoted as:

$$\mathbf{X}_2^{(d,\rho)} = \left\{\mathbf{x}_j \middle| \mathbf{x}_j \in \mathbf{X}_2 \wedge rank(\mathbf{x}_j) \leqslant n_2^{le} \right\}, \tag{16}$$

where $rank(\mathbf{x}_j)$ returns the order of $\mathbf{x}_j$ in $\acute{\mathbf{X}}_2$. The classification of these instances are rectified by label (i.e., $f_2(\cdot)$). To deepen our understanding of the characteristics of the instance trisection, we examine some properties as follows.

Property 1. *An arbitrary instance $\mathbf{x}_j \in \mathbf{X}_2$ will not be selected to conduct label enhancement if $u_{jg} = 0, \forall g = 1, 2, \ldots, m$.*

Proof. According to Equation (15), the n_2^{le} is pertinent to the instance count with local uncertain-prone instances. An instance $\mathbf{x}_j \in \mathbf{X}_2$ is regarded as certain instances if $u_{jg} = 0, \forall g = 1, 2, \ldots, m$ holds. Since $n_2^{le} > 0$, such instance will not be employed for the training of $f_2(\cdot)$. □

Property 2. *At least an instance $\mathbf{x}_j \in \mathbf{X}_2$ applies $f_2(\cdot)$ for label enhancement as long as there is at least an element in U satisfies $u_{jg} > 0$.*

Proof. According to Equation (15), both $\sum_{j=1}^{n_2} sgn\left(u_{jg}\right) \geqslant 1$ and $sgn\left(\sum_{j=1}^{n_2} u_{jg}\right) = 1$ hold if condition $u_{jg} > 0$ is satisfied. n_2^{le} reaches the minimum when the remaining elements are all zeros. Hence, we have $n_2^{le} \geqslant 1$ holds. Therefore, at least an instance (denoted as $\mathbf{x}_j$) is considered for the training of $f_2(\cdot)$. □

Property 3. *Instances in $\mathbf{X}_2$ with $u_{jg} > 0$ do not necessarily apply $f_2(\cdot)$ for label enhancement.*

Proof. According to Equation (14), we select the instances with the largest n_2^{le} global uncertainty for label enhancement with $f_2(\cdot)$. Equation (15) shows that $n_2^{le} \leqslant s$ holds. Therefore, instances in $\mathbf{X}_2$ may not be considered for label enhancement if they are not ranked in the top n_2^{le} of $\acute{\mathbf{X}}_2$. □

3.3. Acting: selective label enhancement

For certain instances identified by the instance trisecting module, we take a straightforward action to leverage the results classified by $f_1(\cdot)$. Label enhancement presents a more comprehensive description of each label, which completes by exploring the credible annotations for each label. Intuitively, the smaller the distance from instance $\mathbf{x}_i$ to hyperplane is, the greater the possibility of misclass the instance. Herein we apply this notion on each label independently and estimate the margin on each arbitrary label l_c as the difference between real-valued outputs and calibrated threshold. Therefore, we take the relative distance definition in Equation (6) to identify the instances with a certain classification on label l_c.

Then we stipulate instances with trustworthy positive pseudo-labels at the label level. For simplicity, we define them as those with bigger than the average relative distance of the positive class pseudo-labels to the classification hyperplane, denoted as:

$$\neg\mathbf{X}_1^{c+} = \left\{\mathbf{x}_i \left|\mathbf{x}_i \in \mathbf{X}_1 \wedge \hat{y}_{ic} = 1 \wedge d_{ic} \geqslant R_c^+ \right.\right\}, \tag{17}$$

where R_c^+ denotes the average relative distance of training instances with the positive pseudo-label class on label l_c to the hyperplane. Similarly, we have instances with trustworthy negative pseudo-labels at the label level, denoted as:

$$\neg\mathbf{X}_1^{c-} = \left\{\mathbf{x}_i \left|\mathbf{x}_i \in \mathbf{X}_1 \wedge \hat{y}_{ic} = 0 \wedge d_{ic} \leqslant R_c^- \right.\right\}, \tag{18}$$

where R_c^- denotes the average relative distance from training instances with the negative pseudo-label class on label l_c to the hyperplane. The instances with trustworthy pseudo-label for label level on l_c are the union of the instances with trustworthy positive pseudo-labels and the instances with trustworthy negative pseudo-labels, denoted as:

$$\neg\mathbf{X}_1^{c} = \neg\mathbf{X}_1^{c+} \cup \neg\mathbf{X}_1^{c-}. \tag{19}$$

We assume that the accurate classification of different labels is of equal importance. Consequently, we take a straightforward strategy to merge all the label-level trustworthy instances, denoted as:

$$\neg\mathbf{X}_1^{(d)} = \bigcup \neg\mathbf{X}_1^{c}, \tag{20}$$

where $c = 1, 2, \cdots, m$. These credible annotations have more salient features. Thus, it is more conducive to discriminating the labels by employing label enhancement on $\neg\mathbf{X}_1^{(d)}$.

3.4. Outcome: evaluation metrics

The outcome part evaluates the effectiveness of trisecting and acting on instances. As an extension of single-label, the evaluation on multi-label includes perspectives of both label and instance. For classification performance, we consider five metrics [61] including *Hamming Loss*, *Ranking Loss*, *One Error*, *Coverage* and *Average Precision*. Let Y_i and $\overline{Y_i}$ denote the relevant and irrelevant label set in ground-truth, n_2 be the unseen instances count, then the equations of metrics are enumerated as:

(1) *Hamming Loss*(abbreviated as Hl): It evaluates the average difference between predictions and ground truth (see Equation (21)). The smaller the value of *Hamming Loss* is, the better the performance of an algorithm becomes.

$$\mathrm{Hl} = \frac{1}{n_2}\sum_{i=1}^{n_2}\frac{1}{l}\left|f\left(\mathbf{x}_i\right)\Delta Y_i\right|, \tag{21}$$

where Δ denotes the set symmetric difference and $|\cdot|$ denotes the set cardinality.

(2) *Ranking Loss*(abbreviated as Rkl): It evaluates the fraction an irrelevant label ranks before the relevant label in label predictions (see Equation (22)). The smaller the value of the *Ranking loss* is, the better the performance of an algorithm becomes.

$$\mathrm{Rkl} = \frac{1}{n_2}\sum_{i=1}^{n_2}\frac{\left|\left\{\left(l_a, l_b\right)\left|r_i\left(l_a\right) > r_i\left(l_b\right) \wedge \left(l_a, l_b\right) \in Y_i \times \overline{Y_i}\right.\right\}\right|}{\left|Y_i\right|\left|\overline{Y_i}\right|}, \tag{22}$$

where $r_i\left(l_j\right)$ denotes the ranking position in ascending order for j-th label on the i-th instance. $|\cdot|$ denotes the set cardinality.

(3) *One Error*(abbreviated as Oe): It evaluates the average fraction that label ranking first in prediction is the irrelevant label (see Equation (23)). The smaller the value of *One Error* is, the better the performance of an algorithm becomes.

$$\mathrm{Oe} = \frac{1}{n_2}\sum_{i=1}^{n_2}\left[\left(\underset{l_j}{\arg\min}\, r_i\left(l_j\right)\right) \notin Y_i\right], \tag{23}$$

where $[\cdot]$ equals to 1 if the condition holds, and equals to 0 otherwise. The operator $r_i\left(l_j\right)$ denotes the ranking position in ascending order for j-th label on the i-th instance.

(4) *Coverage*(abbreviated as Cvg): It evaluates the average fraction for all ground-truth labels in the ranking of label predictions (see Equation (24)). The smaller the *Coverage* is, the better the performance of an algorithm becomes.

$$\mathrm{Cvg} = \frac{1}{n_2}\sum_{i=1}^{n_2}\max_{l_j \in Y_i} r_i\left(l_j\right) - 1, \tag{24}$$

where $r_i\left(l_j\right)$ denotes the ranking position in ascending order for j-th label on the i-th instance.

(5) *Average Precision*(abbreviated as Ap): It evaluates the average precision of actually relevant labels ranking before a relevant label by label predictions (see Equation (25)). The larger the value of *Average Precision* is, the better the performance of an algorithm becomes.

$$\mathrm{Ap} = \frac{1}{n_2}\sum_{i=1}^{n_2}\frac{1}{|Y_i|}\sum_{l_j \in Y_i}\frac{\left|\left\{l_s \in Y_i \left| r_i\left(l_s\right) \leqslant r_i\left(l_j\right)\right.\right\}\right|}{r_i\left(l_j\right)}, \tag{25}$$

where $|\cdot|$ denotes the set cardinality.

3.5. Complexity analysis

We summarize the procedures of instance trisections and 3WDLE in Algorithm 1. Steps 1-12 and 13-24 correspond to the training and testing of 3WDLE, respectively. The computation of kNN is the most time-consuming step. Let $\acute{n}_1$ and $\acute{n}_2$ be the average count of local uncertain-prone instances for $\mathbf{X}_1$ and $\mathbf{X}_2$, L be the average length of label-specific features, and m be the label count. The generation for kNN requires $O\left(\acute{n}_1{}^2 Lm\right)$ and $O\left(\acute{n}_2{}^2 Lm\right)$ for training and testing, respectively.

Algorithm 1: 3WDLE.

```
    Input: $\mathbf{X}_1$, $\mathbf{X}_2$, $\mathbf{Y}_1$, and $k$.
    Output: unseen labels $\hat{\mathbf{Y}}_2^*$.
 1  Construct $f_1(\cdot)$ by using $\mathbf{X}_1$ and $\mathbf{Y}_1$, as described in (1).
 2  Generate pseudo labels $\hat{\mathbf{Y}}_1$ given $\mathbf{X}_1$ as described in (2).
 3  for $c = 1$ to $m$ do
 4      for $i = 1$ to $n_1$ do
 5          Compute $d_{ic}$ as described in Equation (6).
 6          Compute $\neg\mathbf{X}_1^{c+}$ as described in Equation (17).
 7          Compute $\neg\mathbf{X}_1^{c-}$ as described in Equation (18).
 8          Generate $\neg\mathbf{X}_1^{c}$ as described in Equation (19).
 9      end
10  end
11  Generate $\neg\mathbf{X}_1^{(d)}$ as described in Equation (20).
12  Construct $f_2(\cdot)$ by using $\neg\mathbf{X}_1^{(d)}$ and $\neg\mathbf{Y}_1^{(d)}$ as described in (3).
13  for $c = 1$ to $m$ do
14      for $j = 1$ to $n_2$ do
15          Compute $d_{jc}$ as described in Equation (6).
16          Compute $\mathbf{X}_2^{c+}$ as described in Equation (7)
17          Compute $\mathbf{X}_2^{c-}$ as described in Equation (8).
18          Compute $\rho_{jc}$ as described in Equation (9).
19          Compute $u_{jc}$ as described in Equation (10).
20      end
21  end
22  Generate $\mathbf{X}_2^{(d_c,\rho_c)}$ as described in Equation (12).
23  Generate $\acute{\mathbf{X}}_2$ as described in Equation (14).
24  Generate $\mathbf{X}_2^{(d,\rho)}$ as described in Equation (16).
25  Generate $\hat{\mathbf{Y}}_2^*$ as described in Equation (5).
```

4. Experiments

4.1. Dataset characteristics

To demonstrate the effectiveness and efficiency of the proposed model, we compare classification performance on eight multi-label benchmarks from Mulan,[1] Meka[2] and MDDM.[3] In Table 1, for each dataset, "# Instances" means the number of

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

[2] http://waikato.github.io/meka/datasets/.

[3] http://www.lamda.nju.edu.cn/code_MDDM.ashx.

Table 1
Characteristics of Data Sets.

Data set	# Instances	# Features	# Labels	# Cardinality	Source
art	5000	462	26	1.64	MDDM
bibtex	7395	1836	159	2.402	Mulan
business	5000	438	32	1.59	MDDM
enron	1702	1001	53	3.378	Meka
languagelog	1460	1004	75	1.18	Meka
medical	978	1449	45	1.245	Mulan
scene	2407	294	6	1.074	Mulan

instances, “# Features” means the number of features, “# Labels” means the total number of class labels, and “# Cardinality” means the average number of labels per instance of a dataset.

4.2. Experimental settings

We conduct comparisons in two groups. In the first group, we explore whether label enhancement under the framework of three-way decisions is conducive to boosting classification performance. For this sake, we compare 3WDLE with ML*k*NN, LIFT, MLTSVM, Glocal, HNOML, and fRA*k*EL. Detailed settings are as follows.

- 3WDLE: Proposed method. $f_1(\cdot)$ and $f_2(\cdot)$ are implemented via LLSF and LEMLL, respectively. The neighbor size k takes the empirical value of 10.
- LLSF[4] [10]: This method learns a label-specific feature representation for all labels based on logical labels. Parameters of δ, η are tuned in $\{2^{-10}, 2^{-9}, \dots, 2^{9}, 2^{10}\}$. The calibrated threshold τ_1 is fixed at 0.5.
- ML*k*NN[5] [12]: It learns a conditional probability distribution on all features within the adapted k-neighborhood. The value k takes the empirical value of 10.
- LIFT[6] [53]: It learns different feature representations to determine label association. The ratio parameter is searched in $\{0.1, 0.2, \dots, 0.5\}$.
- MLTSVM[7] [14]: It learns distance differences based on multiple nonparallel hyperplanes. The penalty 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[8] [16]: It learns a mapping from feature space to latent labels via low-rank decomposition. The penalty takes the empirical value 1.
- HNOML [18]: It enriches the label correlation by utilizing label embedding. Penalty parameter α, β, and γ are searched in $\{10^{-3}, 10^{-2}, \dots, 10^{3}\}$.
- fRA*k*EL[9] [60]: A fast version of Random k-label sets. The label set size takes the empirical value of 3, and we configure the base classifier count as twice the label cardinality count. The base classifier adopts LibLinear.[10]

The entire experiments are implemented using Matlab R2017b under a desktop PC with Intel(R) Core i7 processor (2.60GHz) and 8GB RAM. All parameters are selected via five-fold cross-validation on the training set.

4.3. Results

We evaluate the classification performance for considered algorithms on five evaluation metrics and report them in Table 2, Table 3, Table 4, Table 5 and Table 6. The lower ranking of 3WDLE than LLSF on all comparisons illustrates the effectiveness of label enhancement and achieves satisfactory performance over a collection of state-of-the-art algorithms. For comprehension, we present the algorithm ranking on the metric view (i.e., average ranking) and the dataset view (followed by mean ± standard deviation). From the measure metric view, 3WDLE ranks first at 60% cases, second at 20% and third at only 20%. From the dataset view, 3WDLE ranks first in 40% ($\frac{14}{35}$), second at 22.9% ($\frac{4}{35}$), third at 28.6% ($\frac{10}{35}$), and in the second half at only 20% ($\frac{7}{35}$). It receives the best performance on metric *Coverage* (with first place at 100%) and worst on metric *One Error* (with the second half at 28.6%).

Friedman test [62] is employed to calculate the relative performance among multiple algorithms over selected datasets. Given k comparing algorithms and N datasets, let $R_j = (1/N)\sum_{i=1}^{N} r_i^j$ denote the average rank for the j-th algorithm. With the null hypothesis (i.e., H_0) that all algorithms obtain identical performance, the Friedman statistic F_F is distributed

[4] Code available at https://jiunhwang.github.io/.

[5] Code available at http://www.lamda.nju.edu.cn/code_ML*k*NN.ashx.

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

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

[8] Code available at http://www.lamda.nju.edu.cn/code_Glocal.ashx.

[9] Code available at http://github.com/KKimura360/fast_RAkEL_matlab.

[10] https://www.csie.ntu.edu.tw/~cjlin/liblinear/.

Table 2
Comparisons (mean ± std) on metric *Hamming Loss*. ↓ means the smaller the value is, the better the performance becomes. Numbers in brackets refer to the performance ranking, whereas the Avg rank denotes the average ranking across all benchmarks. Results with the best performance are in the bold size.

Data set	Hamming Loss (↓)							
	3WDLE	LLSF	ML*k*NN	LIFT	MLTSVM	Glocal	HNOML	fRA*k*EL
art	**0.052±0.001**(1)	0.054±0.001(3)	0.060±0.001(5)	0.053±0.001(2)	0.057±0.002(4)	0.066±0.001(7)	0.062±0.001(6)	0.136±0.004(8)
bibtex	**0.012±0.000**(1)	0.015±0.001(5.5)	0.016±0.001(7)	**0.013±0.002**(2.5)	0.017±0.001(8)	0.014±0.001(4)	0.015±0.001(5.5)	0.013±0.001(2.5)
business	0.027±0.001(2)	0.033±0.002(7)	0.028±0.001(3.5)	0.028±0.001(3.5)	**0.025±0.001**(1)	0.029±0.001(5)	0.030±0.001(6)	0.046±0.002(8)
enron	0.047±0.002(4)	0.053±0.002(6)	0.051±0.002(5)	0.046±0.001(2.5)	0.062±0.002(7)	0.076±0.006(8)	0.046±0.001(2.5)	**0.045±0.001**(1)
languagelog	**0.015±0.001**(1)	0.018±0.001(5.5)	0.016±0.001(3)	0.029±0.001(8)	0.018±0.001(5.5)	0.016±0.001(3)	0.016±0.001(3)	0.025±0.001(7)
medical	**0.012±0.001**(1)	0.014±0.001(4.5)	0.015±0.001(7)	0.013±0.001(2)	0.014±0.002(4.5)	0.019±0.003(8)	0.014±0.042(4.5)	0.014±0.001(4.5)
scene	0.102±0.005(4)	0.108±0.004(5)	0.092±0.006(2)	**0.079±0.005**(1)	0.143±0.003(7)	0.111±0.006(6)	0.147±0.011(8)	0.093±0.004(3)
Avg rank	2.000(1)	5.214(6)	4.643(3)	3.071(2)	5.286(7)	5.857(8)	5.071(5)	4.857(4)

Table 3
Comparisons (mean ± std) on metric *Ranking Loss*. ↓ means the smaller the value is, the better the performance becomes. Numbers in brackets refer to the performance ranking, whereas the Avg rank denotes the average ranking across all benchmarks. Results with the best performance are in bold size.

Data set	Ranking Loss (↓)							
	3WDLE	LLSF	ML*k*NN	LIFT	MLTSVM	Glocal	HNOML	fRA*k*EL
art	0.148±0.004(3)	0.152±0.005(5)	0.150±0.007(4)	**0.113±0.005**(1)	0.625±0.006(8)	0.154±0.016(6)	0.122±0.003(2)	0.454±0.015(7)
bibtex	0.088±0.005(2)	0.096±0.004(3)	0.205±0.004(6)	**0.074±0.004**(1)	0.660±0.006(8)	0.160±0.004(5)	0.139±0.005(4)	0.233±0.010(7)
business	0.048±0.002(4)	0.054±0.003(6)	0.038±0.003(2)	**0.031±0.002**(1)	0.250±0.004(8)	0.049±0.007(5)	0.046±0.004(3)	0.185±0.008(7)
enron	0.081±0.006(2)	0.119±0.008(4)	0.092±0.003(3)	**0.077±0.006**(1)	0.499±0.012(6)	0.132±0.008(5)	0.667±0.014(7)	0.679±0.019(8)
languagelog	0.168±0.013(2)	0.193±0.018(4.5)	**0.127±0.005**(1)	0.183±0.023(3)	0.731±0.014(8)	0.193±0.006(4.5)	0.288±0.018(6)	0.555±0.020(7)
medical	0.025±0.005(2)	0.037±0.001(4)	0.043±0.007(6)	0.029±0.006(3)	0.168±0.018(7)	0.038±0.006(5)	**0.021±0.059**(1)	0.206±0.022(8)
scene	0.093±0.013(3)	0.101±0.002(5)	0.085±0.008(2)	**0.064±0.007**(1)	0.278±0.009(8)	0.096±0.005(4)	0.110±0.010(6)	0.155±0.024(7)
Avg rank	2.571(2)	4.500(5)	3.429(3)	1.571(1)	7.571(8)	4.929(6)	4.143(4)	7.286(7)

Table 4
Comparisons (mean ± std) on metric *One Error*. ↓ means the smaller the value is, the better the performance becomes. Numbers in brackets refer to the performance ranking, whereas the Avg rank denotes the average ranking across all benchmarks. Results with the best performance are in bold size.

Data set	One Error (↓)							
	3WDLE	LLSF	ML*k*NN	LIFT	MLTSVM	Glocal	HNOML	fRA*k*EL
art	0.417±0.010(2)	0.474±0.009(5)	0.622±0.007(8)	0.458±0.003(3)	0.523±0.007(7)	0.468±0.020(4)	0.475±0.011(6)	**0.303±0.017**(1)
bibtex	0.357±0.009(2)	0.376±0.008(3)	0.588±0.008(7)	0.386±0.010(4)	0.424±0.013(5)	0.525±0.135(6)	0.618±0.005(8)	**0.195±0.008**(1)
business	0.106±0.010(3.5)	0.111±0.009(5)	0.113±0.007(6)	0.106±0.007(3.5)	0.092±0.006(2)	0.119±0.009(7)	0.126±0.005(8)	**0.068±0.007**(1)
enron	0.219±0.010(2)	0.286±0.025(5)	0.255±0.019(4)	0.240±0.021(3)	**0.139±0.007**(1)	0.293±0.037(6)	0.950±0.017(7)	0.955±0.021(8)
languagelog	0.740±0.018(5)	0.756±0.024(6)	0.722±0.015(4)	0.676±0.013(2)	0.707±0.026(3)	0.983±0.011(8)	0.773±0.009(7)	**0.401±0.028**(1)
medical	0.142±0.027(3)	0.189±0.026(7)	0.235±0.010(8)	0.165±0.014(6)	0.113±0.020(2)	0.157±0.020(5)	0.154±0.177(4)	**0.062±0.016**(1)
scene	0.251±0.027(5)	0.268±0.014(7)	0.243±0.171(4)	0.194±0.021(3)	0.177±0.013(2)	0.264±0.011(6)	0.292±0.014(8)	**0.061±0.015**(1)
Avg rank	3.214(3)	5.429(5)	5.857(6)	3.500(4)	3.143(2)	6.000(7)	6.857(8)	2.000(1)

Table 5
Comparisons (mean ± std) on metric *Coverage*. ↓ means the smaller the value is, the better the performance becomes. Numbers in brackets refer to the performance ranking, whereas the Avg rank denotes the average ranking across all benchmarks. Results with the best performance are in bold size.

Data set	Coverage (↓)							
	3WDLE	LLSF	ML*k*NN	LIFT	MLTSVM	Glocal	HNOML	fRA*k*EL
art	**0.008±0.001**(1)	0.009±0.001(2)	5.369±0.242(6)	4.478±0.181(5)	11.68±0.129(8)	5.903±0.404(7)	0.180±0.005(3)	0.329±0.013(4)
bibtex	**0.001±0.000**(1.5)	**0.001±0.001**(1.5)	0.333±0.004(4)	22.14±1.176(6)	0.503±0.005(5)	37.60±1.113(8)	0.211±0.006(3)	25.52±1.356(7)
business	**0.003±0.000**(1)	0.004±0.001(2)	2.189±0.107(6)	1.949±0.122(5)	9.307±0.296(8)	2.808±0.289(7)	0.087±0.009(3)	0.112±0.005(4)
enron	**0.004±0.001**(1)	0.006±0.001(2)	0.246±0.008(3)	11.94±0.721(6)	30.89±0.774(8)	18.04±0.872(7)	0.831±0.019(4)	0.835±0.040(5)
languagelog	**0.002±0.000**(1)	0.003±0.001(2)	0.159±0.008(3)	13.39±1.557(4)	30.79±1.403(8)	17.86±0.308(6)	27.56±1.346(7)	14.49±0.848(5)
medical	**0.001±0.000**(1.5)	0.001±0.001(1.5)	0.059±0.007(3)	1.954±0.342(5)	4.648±0.700(7)	2.397±0.286(6)	8.123±1.268(8)	1.697±0.426(4)
scene	**0.015±0.002**(1)	0.017±0.002(2)	0.085±0.006(3)	0.395±0.041(5)	0.887±0.052(8)	0.568±0.033(7)	0.230±0.065(4)	0.551±0.050(6)
Avg rank	1.143(1)	1.857(2)	4.000(3)	5.143(6)	7.429(8)	6.857(7)	4.571(4)	5.000(5)

Table 6
Comparisons (mean ± std) on metric *Average Precision*. ↑ means the larger the value is, the better the performance. Numbers in brackets refer to the performance ranking, whereas the Avg rank denotes the average ranking across all benchmarks. Results with the best performance are in bold size.

Data set	Average Precision (↑)							
	3WDLE	LLSF	ML*k*NN	LIFT	MLTSVM	Glocal	HNOML	fRA*k*EL
art	0.617±0.003(3)	0.606±0.005(4.5)	0.519±0.008(7)	**0.622±0.016**(1)	0.461±0.007(8)	0.606±0.021(4.5)	0.619±0.008(2)	0.538±0.015(6)
bibtex	**0.590±0.008**(1)	0.573±0.004(2)	0.360±0.007(6)	0.561±0.006(3)	0.326±0.012(8)	0.358±0.008(7)	0.362±0.007(5)	0.422±0.008(4)
business	0.879±0.007(3)	0.859±0.007(6)	0.882±0.008(2)	**0.894±0.004**(1)	0.756±0.003(7)	0.875±0.011(5)	0.876±0.005(4)	0.708±0.024(8)
enron	**0.704±0.016**(1)	0.644±0.015(4)	0.659±0.007(3)	0.695±0.014(2)	0.450±0.011(6)	0.642±0.013(5)	0.063±0.002(7)	0.059±0.004(8)
languagelog	0.329±0.018(3)	0.319±0.024(4)	0.304±0.009(5)	0.270±0.020(6)	0.267±0.011(7)	**0.387±0.019**(1)	0.245±0.010(8)	0.371±0.020(2)
medical	**0.893±0.014**(1)	0.856 ±0.014(6)	0.816±0.009(7)	0.870±0.012(5)	0.799±0.024(8)	0.873±0.012(4)	0.887±0.151(2)	0.882±0.015(3)
scene	0.846±0.017(3)	0.835±0.006(7)	0.855±0.011(2)	**0.886±0.013**(1)	0.756±0.005(8)	0.840±0.006(6)	0.845±0.008(4)	0.844±0.011(5)
Avg rank	2.143(1)	4.786(6)	4.571(3.5)	2.714(2)	7.429(8)	4.643(5)	4.571(3.5)	5.143(7)

Table 7
Summary of the Friedman Statistics F_F ($k = 8$, $N = 7$) and the critical value at significance level $\alpha = 0.05$ in terms of each evaluation metric (k:# comparing algorithms; N:# data sets).

Metrics	F_F	Critical value
Hamming Loss	13.6905	
Ranking Loss	36.1071	
One Error	24.7976	2.2371
Coverage	38.8571	
Average Precision	20.8214	

according to the F-distribution with $k-1$ degree of freedom as the numerator and $(k-1)(T-1)$ degrees of freedom as the denominator, denoted as:

$$F_F = \frac{(N-1)\,\chi_F^2}{N\,(k-1)-\chi_F^2} \tag{26}$$

where

$$\chi_F^2 = \frac{12N}{k\,(k+1)}\left[\sum_j R_j^2 - \frac{k\,(k+1)^2}{4}\right]. \tag{27}$$

Table 7 presents the Friedman statistics F_F on all metrics and the corresponding critical value in this setting. H_0 on a metric will be rejected only if the corresponding F_F exceeds the critical value, As shown in Table 7, at the significance level $\alpha = 0.05$, with critical value 2.2371, H_0 should be rejected for all evaluation metrics.

Furthermore, by regarding 3WDLE as the control algorithm, we employ Holm procedure [62] to explore whether 3WDLE gains significant superiority against each of the considered algorithms. Without losing generality, we nominate A_1 as 3WDLE. For the other $k-1$ comparing algorithms (i.e., $A_j\,(2 \leqslant j \leqslant k)$), we stipulate A_j as the one which has the $j-1$-th largest average rank across all datasets on a specific evaluation metric. Consequently, we have the test statistic for comparing A_1 (i.e., 3WDLE) with A_j as:

$$z_j = \left(R_1 - R_j\right) \Big/ \sqrt{\frac{k\,(k+1)}{6N}}. \tag{28}$$

Let p_j $(2 \leqslant j \leqslant k)$ denote the p-value of z_j under normal distribution. Given significant level $\alpha = 0.05$, Holm procedure works in a stepwise manner by validating whether the statistics p_j is smaller than $\alpha/(k-j+1)$ in ascending order of j. Specifically, Holm procedure continues until there exists j^*-th step, where j^* denotes the first j such that $p_j \geqslant \alpha/(k-j+1)$ holds.[11] This means 3WDLE is deemed to be significantly different against algorithm A_j, for $j \in \{2, \dots, j^*-1\}$.

It shows from Table 8 that 3WDLE statistically outperforms Glocal on metrics *Hamming Loss* and *Coverage* and statistically outperforms MLTSVM on metrics *Ranking Loss*, *Coverage* and *Average Precision*. 3WDLE seems to be most dominant on *Coverage*, which is statistically superior to overall algorithms except ML*k*NN and LLSF.

5. Discussions

In this section, we discuss more findings regarding 3WDLE. It is the first effort to conduct label enhancement on uncertainty instances for multi-label classification to the best of our knowledge. Our goal is to diminish the uncertainty predictions for unseen instances, and a model with enriched labels devises an effective strategy to construct the uncertainty instances assemble. Such processing simulates the way humans handle uncertainty in classification problems. In other words, we may not require strong evidence if we believe the classification is plausible, whereas seeking a novel perspective for decision-making otherwise. The 3WDLE extends the TAO model to the multi-label classification problem. In particular, the trisecting and acting determine with positive label association, negative label association, and undetermined with label enhancement. However, the classification performance on five metrics corresponds to the outcome. Empirical studies show that the presented combination is a competitive solution for multi-label classification.

The neighbor size k is the only additional parameter in 3WDLE if we regard LLSF and LEMLL as two independent modules. It keeps in line with the recommended setting in ML-*k*NN [12]. The sensitivity of the parameter k is not significant on all metrics across all benchmarks, as illustrated in Fig. 4. Such observations guarantee the minimum artificial impacts.

There is some recent progress on the acceleration of label enhancement like FLE [63]. However, an in-depth review suggests that the acceleration only works on the transformation from logical to numerical labels, whereas the proposed label distribution learning is two-staged. It means FLE only improves the pre-processing of label distribution learning and

[11] If $p_j < \alpha\,/\,(k-j+1)$ holds for all j, j^* takes the value of $k+1$.

Table 8
Comparison of 3WDLE (control algorithm) against the remaining approaches. The test statistics z_i and p-value are determined by the Holm test at significance level α=0.05. Algorithms that are statistically inferior to 3WDLE are shown in bold size.

Hamming Loss				
j	algorithm	z_j	*p*	Holm
2	**Glocal**	-2.945942	0.0032	0.007
3	MLTSVM	-2.509506	0.0121	0.008
4	LLSF	-2.454951	0.0141	0.010
5	HNOML	-2.345842	0.0190	0.013
6	fRA*k*EL	-2.182179	0.0291	0.017
7	ML*k*NN	-2.018515	0.0435	0.025
8	LIFT	-0.818317	0.4132	0.050
Ranking Loss				
j	algorithm	z_j	*p*	Holm
2	**MLTSVM**	-3.8188	1.34e-04	0.007
3	**fRA*k*EL**	-3.6012	0.0003	0.008
4	Glocal	-1.8010	0.0717	0.010
5	LLSF	-1.4733	0.1407	0.013
6	HNOML	-1.2006	0.2299	0.017
7	ML*k*NN	-0.6553	0.5123	0.025
8	LIFT	0.7638	1.0000	0.050
One Error				
j	algorithm	z_j	*p*	Holm
2	**HNOML**	-2.7828	0.0054	0.007
3	Glocal	-2.1283	0.0333	0.008
4	ML*k*NN	-2.0191	0.0435	0.010
5	LLSF	-1.6921	0.0906	0.013
6	LIFT	-0.2185	0.8270	0.017
7	MLTSVM	0.0542	1.0000	0.025
8	fRA*k*EL	0.9274	1.0000	0.050
Coverage				
j	algorithm	z_j	*p*	Holm
2	**MLTSVM**	-4.800794	2e-06	0.007
3	**Glocal**	-4.364358	1.3e-05	0.008
4	**LIFT**	-3.055050	0.0023	0.010
5	**fRA*k*EL**	-2.945942	0.0032	0.013
6	HNOML	-2.618615	0.0088	0.017
7	ML*k*NN	-2.182179	0.0291	0.025
8	LLSF	-0.545545	0.5854	0.05
Average Precision				
j	algorithm	z_j	*p*	Holm
2	**MLTSVM**	-4.037031	5.4e-05	0.007
3	fRA*k*EL	-2.291288	0.0219	0.008
4	LLSF	-2.018515	0.0435	0.010
5	Glocal	-1.909407	0.0562	0.013
6	ML*k*NN	1.854852	0.0636	0.017
7	HNOML	-1.854852	0.0636	0.025
8	LIFT	-0.436436	0.6625	0.050

still requires another label distribution learning model to complete the multi-label classification. LEMLL is an end-to-end schema which combines label enhancement and numerical learning in a unified framework. Thus, we employ LEMLL instead of FLE to do label enhancement learning latent label importance.

6. Conclusions

Label enhancement raises the upper bound of classification accuracy since it offers more discriminative descriptions of the multi-label semantics. By automatically conducting the label enhancement, the significant reduction in label ambiguity does not require considerable cost in large-scale delicate annotations. However, the uncertainty of logical labels research has not been examined. In this paper, we proposed a novel model called 3WDLE to deal with the label ambiguity of multi-label by selectively enhancing logical labels under the framework of three-way decisions. It is different from conventional multi-label learning algorithms, which learn a model on labels represented by single granules (i.e., either logical or numerical labels). With label-specific features and pseudo-labels, we devise instance selection principles in a bottom-up manner. Label

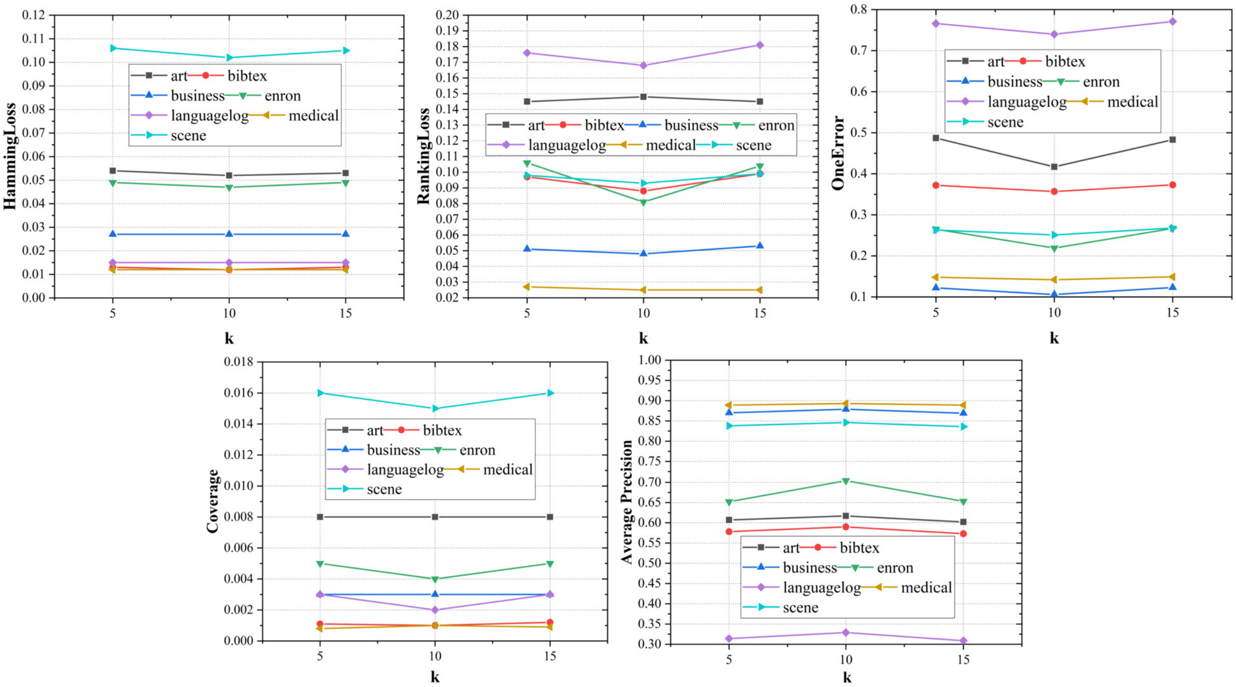

Fig. 4. Performance fluctuations on the varying count of nearest neighborhood.

enhancement is selectively employed to reduce classification uncertainty. By replacing enhanced results, we demonstrate that 3WDLE significantly improves classification performance.

In the future, we will examine more combination methods of label-specific algorithms and label enhancement to seek various guidelines. Meanwhile, we will develop an advanced instance selection principle by resorting to optimization theory.

CRediT authorship contribution statement

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

Acknowledgements

The authors thank the anonymous reviewers for their constructive comments and valuable suggestions. This work is supported by China Postdoctoral Science Foundation (Grant No. 2022M713491), the National Natural Science Foundation of China (Grant No. 61976158, 61976160, 62076182, 62163016, 62006172, 61906137) and is also partially supported by the Jiangxi “Double Thousand Plan” (Grant No. 20212ACB202001).

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] E. Gibaja, S. Ventura, Multi-label learning: a review of the state of the art and ongoing research, Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 4 (6) (2014) 411–444.
[3] E. Gibaja, S. Ventura, A tutorial on multilabel learning, ACM Comput. Surv. 47 (3) (2015) 1–38.
[4] W.W. Liu, X.B. Shen, H.B. Wang, I.W. Tsang, The emerging trends of multi-label learning, IEEE Trans. Pattern Anal. Mach. Intell. (2021), https://doi.org/10.1109/TPAMI.2021.3119334.
[5] S.M. Tabatabaei, S. Dick, W.S. Xu, Toward non-intrusive load monitoring via multi-label classification, IEEE Trans. Smart Grid 8 (1) (2017) 26–40.
[6] H.Z. Fu, J. Cheng, Y.W. Xu, D.W.K. Wong, J. Liu, X.C. Cao, Joint optic disc and cup segmentation based on multi-label deep network and polar transformation, IEEE Trans. Med. Imaging 37 (7) (2018) 1597–1605.
[7] Y.C. Wei, W. Xia, M. Lin, J.S. Huang, B.B. Ni, J. Dong, Y. Zhao, S.C. Yan, HCP: a flexible cnn framework for multi-label image classification, IEEE Trans. Pattern Anal. Mach. Intell. 38 (9) (2016) 1901–1907.

[8] M.R. Boutell, J. Luo, X. Shen, C.M. Brown, Learning multi-label scene classification, Pattern Recognit. 37 (9) (2004) 1757–1771.

[9] G. Tsoumakas, I. Vlahavas, Random k-labelsets: an ensemble method for multilabel classification, Lect. Notes Artif. Intell. 4701 (2007) 406–417.

[10] J. Huang, G.R. Li, Q.M. Huang, X.D. Wu, Learning label-specific features and class-dependent labels for multi-label classification, IEEE Trans. Knowl. Data Eng. 28 (12) (2016) 3309–3323.

[11] M.L. Zhang, Z.H. Zhou, Multilabel neural networks with applications to functional genomics and text categorization, IEEE Trans. Knowl. Data Eng. 18 (10) (2006) 1479–1493.

[12] M.L. Zhang, Z.H. Zhou, ML-*k*NN: a lazy learning approach to multi-label learning, Pattern Recognit. 40 (7) (2007) 2038–2048.

[13] Q.Y. Wu, M.K. Tan, H.J. Song, J. Chen, M.K. Ng, ML-FOREST: a multi-label tree ensemble method for multi-label classification, IEEE Trans. Knowl. Data Eng. 28 (10) (2016) 2665–2680.

[14] W.J. Chen, Y.H. Shao, C.N. Li, N.Y. Deng, MLTSVM: a novel twin support vector machine to multi-label learning, Pattern Recognit. 52 (2016) 61–74.

[15] Z. Ahmadi, S. Kramer, A label compression method for online multi-label classification, Pattern Recognit. Lett. 111 (2018) 64–71.

[16] 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.

[17] J. Huang, L.C. Xu, J. Wang, L. Feng, K. Yamanishi, Discovering latent class labels for multi-label learning, in: Proc. Twenty-ninth Int. Joint Conf. Artif. Intell., 2020, pp. 3058–3064.

[18] C.Q. Zhang, Z.W. Yu, H.Z. Fu, P.F. Zhu, L. Chen, Q.H. Hu, Hybrid noise-oriented multilabel learning, IEEE Trans. Cybern. 50 (6) (2020) 2837–2850.

[19] S.P. Xu, H.R. Ju, L. Shang, W. Pedrycz, X.B. Yang, C. Li, Label distribution learning: a local collaborative mechanism, Int. J. Approx. Reason. 121 (2020) 59–84.

[20] X.Y. Jia, Z.C. Li, X. Zheng, W.W. Li, S.J. Huang, Label distribution learning with label correlations on local samples, IEEE Trans. Knowl. Data Eng. 33 (4) (2021) 1619–1631.

[21] T. Wen, W.W. Li, L. Chen, X.Y. Jia, Semi-supervised label enhancement via structured semantic extraction, Int. J. Mach. Learn. Cybern. 13 (4) (2021) 1131–1144.

[22] W.B. Qian, Y.S. Xiong, J. Yang, W.H. Shu, Feature selection for label distribution learning via feature similarity and label correlation, Inf. Sci. 582 (2022) 38–59.

[23] X. Geng, Label distribution learning, IEEE Trans. Knowl. Data Eng. 28 (7) (2016) 1734–1748.

[24] A. Tao, N. Xu, X. Geng, Labeling information enhancement for multi-label learning with low-rank subspace, in: Pac. Rim Int. Conf. Arti. Intelli., 2018, pp. 671–683.

[25] Y.K. Li, M.L. Zhang, X. Geng, Leveraging implicit relative labeling importance information for effective multi-label learning, in: IEEE Int. Conf. Data Min., 2015, pp. 251–260.

[26] N. Xu, A. Tao, X. Geng, Label enhancement for label distribution learning, in: Proc. Twenty-seventh Int. J. Conf. Artifi. Intelli., 2018, pp. 2926–2932.

[27] R.F. Shao, N. Xu, X. Geng, Multi-label learning with label enhancement, in: IEEE Conf. Data Min., 2018, pp. 437–446.

[28] Y.Y. Yao, Three-way decision: an interpretation of rules in rough set theory, in: Fourth Int. Conf. Rough Sets and Knowl. Tech., 2009, pp. 642–649.

[29] Y.Y. Yao, The superiority of three-way decisions in probabilistic rough set models, Inf. Sci. 181 (6) (2011) 1080–1096.

[30] Y.Y. Yao, Three-way decision and granular computing, Int. J. Approx. Reason. 103 (2018) 107–123.

[31] Y.Y. Yao, Tri-level thinking: models of three-way decision, Int. J. Mach. Learn. Cybern. 11 (2020) 947–959.

[32] R.S. Ren, L. Wei, The attribute reductions of three-way concept lattices, Knowl.-Based Syst. 99 (2016) 92–102.

[33] X.Y. Zhang, H. Yao, Z.Y. Lv, D.Q. Miao, Class-specific information measures and attribute reducts for hierarchy and systematicness, Inf. Sci. 563 (2021) 196–225.

[34] C. Gao, Z.C. Wang, J. Zhou, Three-way approximate reduct based on information-theoretic measure, Int. J. Approx. Reason. 142 (2022) 324–337.

[35] H.L. Zhi, J.J. Qi, T. Qian, L. Wei, Three-way dual concept analysis, Int. J. Approx. Reason. 114 (2019) 151–165.

[36] K.H. Yuan, W.H. Xu, W.T. Li, W.P. Ding, An incremental learning mechanism for object classification based on progressive fuzzy three-way concept, Inf. Sci. 584 (1) (2022) 127–147.

[37] X.W. Xin, J.H. Song, Z.A. Xue, W.W. Peng, Intuitionistic fuzzy three-way formal concept analysis based attribute correlation degree, J. Intell. Fuzzy Syst. 40 (1) (2021) 1567–1583.

[38] H. Yu, Y. Chen, P. Lingras, G.Y. Wang, A three-way cluster ensemble approach for large-scale data, Int. J. Approx. Reason. 115 (2019) 32–49.

[39] H. Yu, X.C. Wang, G.Y. Wang, X.H. Zeng, An active three-way clustering method via low-rank matrices for multi-view data, Inf. Sci. 507 (2020) 823–839.

[40] Q.P. Shen, Q.H. Zhang, F. Zhao, G.Y. Wang, Adaptive three-way c-means clustering based on the cognition of distance stability, Cogn. Comput. 14 (2) (2022) 563–580.

[41] X.Y. Jia, Z. Deng, F. Min, D. Liu, Three-way decisions based feature fusion for Chinese irony detection, Int. J. Approx. Reason. 113 (2019) 324–335.

[42] G.M. Lang, D.Q. Miao, H. Fujita, Three-way group conflict analysis based on pythagorean fuzzy set theory, IEEE Trans. Fuzzy Syst. 28 (3) (2020) 447–461.

[43] Q.H. Zhang, C.C. Yang, G.Y. Wang, A sequential three-way decision model with intuitionistic fuzzy numbers, IEEE Trans. Syst. Man Cybern. Syst. 51 (5) (2021) 2640–2652.

[44] X.Y. Zhang, H.Y. Gou, Z.Y. Lv, D.Q. Miao, Double-quantitative distance measurement and classification learning based on the tri-level granular structure of neighborhood system, Knowl.-Based Syst. 217 (2021) 106799.

[45] J. Chen, Y. Xu, S. Zhao, Y.P. Zhang, AH3: an adaptive hierarchical feature representation model for three-way decision boundary processing, Int. J. Approx. Reason. 130 (2021) 259–272.

[46] C.M. Jiang, D.D. Guo, L.J. Sun, Effectiveness measure for TAO model of three-way decisions with interval set, J. Intell. Fuzzy Syst. 40 (6) (2021) 11071–11084.

[47] Y.J. Zhang, D.Q. Miao, Z.F. Zhang, J.F. Xu, S. Luo, A three-way selective ensemble model for multi-label classification, Int. J. Approx. Reason. 103 (2018) 394–413.

[48] F.J. Ren, L. Wang, Sentiment analysis of text based on three-way decisions, J. Intell. Fuzzy Syst. 33 (1) (2017) 245–254.

[49] Y.J. Zhang, D.Q. Miao, W. Pedrycz, T.N. Zhao, J.F. Xu, Y. Yu, Granular structure-based incremental updating for multi-label classification, Knowl.-Based Syst. 189 (2020) 105066.

[50] Y.J. Zhang, T.N. Zhao, D.Q. Miao, W. Pedrycz, Granular multilabel batch active learning with pairwise label correlation, IEEE Trans. Syst. Man Cybern. Syst. 52 (5) (2022) 3079–3091.

[51] C.M. Zhu, D.J. Cao, S.P. Guo, R.G. Zhou, Y.L. Dong, D.Q. Miao, Weak-label-based global and local multi-view multi-label learning with three-way clustering, Int. J. Mach. Learn. Cybern. 13 (2021) 1337–1354.

[52] W.B. Qian, J.T. Huang, Y.L. Wang, Y.H. Xie, Label distribution feature selection for multi-label classification with rough set, Int. J. Approx. Reason. 128 (2021) 32–55.

[53] M.L. Zhang, L. Wu, LIFT: multi-label learning with label-specific features, IEEE Trans. Pattern Anal. Mach. Intell. 37 (1) (2015) 107–120.

[54] Y.M. Guo, F.L. Chung, G.Z. Li, J.C. Wang, J.C. Gee, Leveraging label-specific discriminant mapping features for multi-label learning, ACM Trans. Knowl. Discov. Data 13 (2) (2019) 24.

[55] 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.

[56] N. Xu, Y.P. Liu, X. Geng, Label enhancement for label distribution learning, IEEE Trans. Knowl. Data Eng. 33 (4) (2021) 1632–1643.

[57] N. Xu, J. Shu, Y.P. Liu, X. Geng, Variational label enhancement, in: Proc. Int. Conf. Mach. Learn., 2020, pp. 10597–10606.
[58] X.Y. Jia, Y.N. Lu, F.W. Zhang, Label enhancement by maintaining positive and negative label relation, IEEE Trans. Knowl. Data Eng. (2021), https://doi.org/10.1109/TKDE/2021.3093099.
[59] Q.H. Zheng, J.H. Zhu, H.Y. Tang, X.Y. Liu, Z.Y. Li, H.M. Lu, Generalized label enhancement with sample correlations, IEEE Trans. Knowl. Data Eng. (2021), https://doi.org/10.1109/TKDE.2021.3073157.
[60] K. Kimura, M. Kudo, L. Sun, S. Koujaku, Fast random k-labelsets for large-scale multi-label classification, in: Proc. Int. Conf. Pattern Recog., 2017, pp. 438–443.
[61] R. Schapire, Y. Singer, A boosting-based system for text categorization, Mach. Learn. 39 (2/3) (2000) 135–168.
[62] J. Demsar, Statistical comparisons of classifiers over multiple data sets, J. Mach. Learn. Res. 7 (2006) 1–30.
[63] K. Wang, N. Xu, M.G. Ling, X. Geng, Fast label enhancement for label distribution learning, IEEE Trans. Knowl. Data Eng. (2021), https://doi.org/10.1109/TKDE.2021.3092406.