mathematics

Article

Intuitionistic Fuzzy-Based Three-Way Label Enhancement for Multi-Label Classification

Tianna Zhao [1,†], Yuanjian Zhang [2,3,*,†] and Duoqian Miao [1,*]

1 Department of Computer Science and Technology, Tongji University, Shanghai 201804, China; 1810375@tongji.edu.cn
2 China UnionPay Co., Ltd., Shanghai 201201, China
3 Postdoctoral Research Station of Computer Science and Technology, Fudan University, Shanghai 200433, China
* Correspondence: zhangyuanjian@unionpay.com (Y.Z.); dqmiao@tongji.edu.cn (D.M.)
† These authors contributed equally to this work.

Abstract: Multi-label classification deals with the determination of instance-label associations for unseen instances. Although many margin-based approaches are delicately developed, the uncertainty classifications for those with smaller separation margins remain unsolved. The intuitionistic fuzzy set is an effective tool to characterize the concept of uncertainty, yet it has not been examined for multi-label cases. This paper proposed a novel model called intuitionistic fuzzy three-way label enhancement (IFTWLE) for multi-label classification. The IFTWLE combines label enhancement with an intuitionistic fuzzy set under the framework of three-way decisions. For unseen instances, we generated the pseudo-label for label uncertainty evaluation from a logical label-based model. An intuitionistic fuzzy set-based instance selection principle seamlessly bridges logical label learning and numerical label learning. The principle is hierarchically developed. At the label level, membership and non-membership functions are pair-wisely defined to measure the local uncertainty and generate candidate uncertain instances. After upgrading to the instance level, we select instances from the candidates for label enhancement, whereas they remained unchanged for the remaining. To the best of our knowledge, this is the first attempt to combine logical label learning with numerical label learning into a unified framework for minimizing classification uncertainty. Extensive experiments demonstrate that, with the selectively reconstructed label importance, IFTWLE achieves statistically superior over the state-of-the-art multi-label classification algorithms in terms of classification accuracy. The computational complexity of this algorithm is $O(n^2mk)$, where n, m, and k denote the unseen instances count, label count, and average label-specific feature size, respectively.

Keywords: multi-label classification; uncertainty; three-way decisions; intuitionistic fuzzy number; instance selection principle; label enhancement

MSC: 03B52; 68T30; 68T37

Citation: Zhao, T.; Zhang, Y.; Miao, D. Intuitionistic Fuzzy-Based Three-Way Label Enhancement for Multi-Label Classification. *Mathematics* **2022**, *10*, 1847. https://doi.org/10.3390/math10111847

Academic Editors: Shu Zhao and Catalin Stoean

Received: 11 April 2022
Accepted: 24 May 2022
Published: 27 May 2022

1. Introduction

In multi-label settings [1–3], an instance is associated with multiple labels simultaneously. For example, a picture may be relevant to tags, such as *sky*, *ocean*, and *seagull*; a book may cover topics, such as *sports* and *art*. The goal of multi-label classification is to learn mapping from a feature space to a label space, such that the label associations of unseen instances can be determined. It widely applies in domains involving smart grid management [4], disease diagnosis [5], and image classification [6].

Traditionally, multi-label classification schema builds upon logical labels and provides qualitative associations between instances and labels. For robustness, many researchers focus on the extensions of linear or hyperplane-based models, and their work can be roughly categorized as problem transformation and algorithm adaptation, respectively. The previous transforms multi-label classification into a collection of simplified learning scenarios, such as single-label classification. Representative work involves binary relevance [7], random *k*-labelsets [8], and learning label-specific feature (LLSF) [9]. Comparatively, the latter extends the existing algorithms to simultaneously generate multi-outputs. Representative work includes multi-label forest (ML-Forest) [10] and multi-label twin support vector machine (MLTSVM) [11]. However, both strategies suffered from the calibrated threshold problem. Intuitively, the uncertainty of the label association is larger if the output is close to the threshold, and is smaller otherwise. Hence, we need labels with stronger supervision to boost the model performance.

Regarding a numerical label–this quantifies how much information a label has in describing an instance. For example, a facial expression can be interpreted as the combination of *slight sadness*, *some anger*, and *a bit of disgust*. The fitting of the numerical label is defined as label distribution learning [12]. Although numerical labels offer more discriminative information than logical labels, it is expensive to annotate all numerical labels manually. One feasible solution is to learn numerical labels from logical labels, also known as *label enhancement* [13–16]. However, the research studies have indistinguishably employed label enhancements for all instances, regardless of the uncertainty differences.

Three-way decisions [17], also known as trisecting–acting–outcome (TAO) model [18–20], is an emerging decision theory for problem solving with uncertainty [21–23]. It originates from the semantic explanations for three regions induced by rough sets and has become an emerging theory in the soft computing community [24–27]. The sequential three steps are trisecting, acting and outcome characterizing the routines on uncertainty. Typically, the trisecting step divides the information into three non-overlapping regions, with two regions as certain and the others as uncertain; the acting step takes the positive/negative strategy w.r.t. of certain objects whereas takes the deferment strategy w.r.t. of uncertain objects; the outcome step evaluates the performance induced by trisecting and acting. Such routines may continue until the uncertainty is negligible and the classification performance is improved. Existing three-way-based multi-label classifications [28–32] deal with label uncertainty either from ensemble features or ensemble algorithms, whereas the ensemble on logical and numerical labels remains untouched.

1.1. Motivation

The semantic uncertainty analysis on multi-label classification has limitations. The fuzzy set shows the effectiveness in measuring the membership degree towards multi-label [33–35], but it fails to consider the non-membership degree. As a generalization of the fuzzy set, an intuitionistic fuzzy number (IFN) is effective at quantifying the vagueness of a qualitative instance-label assignment [36], which offers heuristic information to select instances for label enhancement. This paper presents an intuitionistic fuzzy-based three-way label enhancement (IFTWLE) model. It implements three-way decisions by "trisecting" unseen instances from label-specific learning and by "acting" label enhancement of uncertain instances. Inspired by empirical studies on single-label [37–39], we employed an intuitionistic fuzzy number to search instances with uncertainty classifications. Concretely speaking, IFTWLE applies IFN at the label level and defines a pair of membership and non-membership functions for every unseen instance based on the generated pseudo-labels. The membership functions evaluate the weighted closeness of instances belonging to the specified class, whereas the non-membership functions evaluate the possibility of instances belonging to the complementary class. The ultimate uncertainty instances are determined via an aggregation function. Consequently, we preserve the predicted labels if the classifications are plausible and exploit the numerical label otherwise.

1.2. Contribution

Compared with the existing multi-label classifications models, our contributions are as follows:

(1) For the first time, cascade learning of logical labels and numerical labels of multi-labels are presented and unified under the semantics of classification uncertainty. Determination of the label association for unseen instances broadens the application of the three-way decisions theory.
(2) A novel instance selection principle in two stages is presented to integrate logical label learning with numerical label learning. In this way, instances that exhibit significant uncertainty across most labels are enhanced with better discrimination (regarding label importance).
(3) This is the first attempt to employ an intuitionistic fuzzy set on multi-label classification. It addresses the issue of identifying potentially uncertain instances on each label without stipulating many hyperparameters.
(4) The proposed IFTWLE has demonstrated effectiveness across many domains. The computational complexity is proportional to the quadratic instances count and linear to the label scale and label-specific features count.

The remainder of the paper is organized as follows. Section 2 reviews some preliminaries on label-specific learning, label enhancement, and the intuitionistic fuzzy set, Section 3 presents our proposed model for multi-label classification; experimental results are reported in Section 4. Section 5 concludes this work and identifies our future directions.

2. Preliminaries

In this section, we briefly review some preliminaries regarding label-specific feature learning and label enhancement.

2.1. Label-Specific Feature Learning

Label-specific feature [9,40–42] assumes that each label has unique characteristics and can be described by different feature subsets. For the simplicity of computation, the learning label-specific feature (LLSF) [9] 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 feature 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 to the feature space.
(3) Shareability: the cardinality of common features of two label-specific features with strong label correlations should be larger than those with weak label correlations.

Based on the previous hypotheses, the objective function is formulated as:

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

where $\mathbf{X}$ and $\mathbf{Y}$ represent the features and labels of multi-label, $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_i, \ldots, \mathbf{w}_m]$, and the basic element $\mathbf{w}_i$ represents the weight for the label-specific feature of l_i, which is composed of the non-zero terms in $\mathbf{w}_i$. $\|\cdot\|_F^2$ denotes the Frobenius norm. $\mathbf{R} = [r_{ij}]$ represents a matrix composed of the 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 trace. 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 the 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:

$$\hat{\mathbf{Y}} = sgn(\mathbf{XW} - \tau) \quad (2)$$

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

2.2. Label Enhancement

Label enhancement assumes that each instance is intrinsically represented by real-valued labels and, thus, can be recovered from qualitative logical label ($\mathbf{Y}$) to quantitative numerical label ($\mathbf{U}$) via the instance-level or label-level smoothness [15,16,43,44]. The distribution of numerical labels describes the relative importance of different labels in describing a given instance.

To guarantee the effectiveness, three hypotheses are presented in label enhancement multi-label learning (LEMLL) [16].

(1) Linear relevance: the mapping from feature space to numerical label $g : \mathbf{X} \to \mathbf{U}$ follows a linear model.
(2) Label similarity: the values of learnt numerical labels should be approximated to the original logical labels.
(3) Topology consistency: the instances with similar features share similar numerical label values.

Based on the previous assumptions, the objective function is given as:

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

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

For predictions on unseen instances, LEMLL leverages a kernel logistic regression, which is denoted as:

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

2.3. Intuitionistic Fuzzy Set

Suppose an arbitrary instance $\mathbf{x}_i \in \mathbf{X}$ has the corresponding label $\mathbf{y}_i$, then for a nonempty set $\mathbf{X}$, the intuitionistic fuzzy set is defined as follows:

$$\widetilde{A} = \{(\mathbf{x}_i, \mu_{\widetilde{A}}(\mathbf{x}_i), \nu_{\widetilde{A}}(\mathbf{x}_i))|\mathbf{x}_i \in X\} \tag{5}$$

where $\mu_{\widetilde{A}}(\mathbf{x}_i)$ is the membership of instance $\mathbf{x}_i$ expressing the chances of instance $\mathbf{x}_i$ being a particular class A, $\nu_{\widetilde{A}}(\mathbf{x}_i)$ is the non-membership of instance $\mathbf{x}_i \in \mathbf{X}$ representing the possibility of instance $\mathbf{x}_i$ not related to class A. μ and ν are functions mapping from $\mathbf{X}$ to [0,1]; the following two conditions are satisfied:

1. $\mu_{\widetilde{A}}(\mathbf{x}_i) \in [0,1], \nu_{\widetilde{A}}(\mathbf{x}_i) \in [0,1]$;
2. $0 \leq \mu_{\widetilde{A}}(\mathbf{x}_i) + \nu_{\widetilde{A}}(\mathbf{x}_i) \leq 1$.

The hesitation of instance $\mathbf{x}_i$ is defined as:

$$\pi_{\widetilde{A}}(\mathbf{x}_i) = 1 - \mu_{\widetilde{A}}(\mathbf{x}_i) - \nu_{\widetilde{A}}(\mathbf{x}_i) \tag{6}$$

which measures the hesitation of instance $\mathbf{x}_i$ related to class A. $\mu_{\widetilde{A}}$ and $\nu_{\widetilde{A}}$ construct the intuitionistic fuzzy number IFN: $\alpha = (\mu_{\widetilde{A}}, \nu_{\widetilde{A}})$, $S(\alpha)$ is used to compare two intuitionistic fuzzy numbers:

$$S(\alpha) = \mu_{\widetilde{A}} - \nu_{\widetilde{A}} \tag{7}$$

which could compare instances $(\mathbf{x}_i, \mathbf{y}_i, \mu_{\widetilde{A}}(\mathbf{x}_i), \nu_{\widetilde{A}}(\mathbf{x}_i))$ with $(\mathbf{x}_j, \mathbf{y}_j, \mu_{\widetilde{A}}(\mathbf{x}_j), \nu_{\widetilde{A}}(\mathbf{x}_j))$, and through the value of $S(\alpha_i)$ and $S(\alpha_j)$, we could determine whether $\mathbf{x}_i$ or $\mathbf{x}_j$ is more likely to be associated with class A.

3. Proposed Model

3.1. Notations

For ease of reference, we present a nomenclature including the major notations and mathematical meanings in Table 1.

Table 1. Notation of IFTWLE.

Notation	Mathematical Meanings
$Card(\cdot)$	set cardinality
$\mathbf{X}_1$	multi-label instances set
$\mathbf{Y}_1$	logical label set
$\mathbf{X}_2$	unseen instances set
$\mathbf{x}_i$	an instance
$\mathbf{y}_i$	logical label set of the instance $\mathbf{x}_i$
$\hat{\mathbf{y}}_i$	pseudo-label set of the instance $\mathbf{x}_i$ learnt by f_1
$\mathbf{u}_i$	numerical label set of the instance $\mathbf{x}_i$
$\mu_c(\mathbf{x}_i)$	membership degree of $\mathbf{x}_i$ on label l_c
$\nu_c(\mathbf{x}_i)$	non-membership degree of $\mathbf{x}_i$ on label l_c
$\pi_c(\mathbf{x}_i)$	hesitation degree of $\mathbf{x}_i$ on label l_c
$\mathbf{X}_2^{(\mu,\nu)}$	uncertain instances set on all labels
$\neg\mathbf{X}_2^{(\mu,\nu)}$	certain instances set on all labels
$f_1(\cdot)$	function of logical label-based learning
$f_2(\cdot)$	function of numerical label-based learning
$\hat{\mathbf{Y}}_2^*$	final predicted multi-label set of $\mathbf{X}_2$
l_c	label c
$\mathbf{Y}_2^{(\mu,\nu)}$	the label set of $\mathbf{X}_2^{(\mu,\nu)}$ learnt by f_1
$\neg\mathbf{Y}_2^{(\mu,\nu)}$	the label set of $\neg\mathbf{X}_2^{(\mu,\nu)}$ learnt by f_1
$\mathbf{X}_2^{(\mu_c,\nu_c)}$	uncertain instance set on label l_c
$\neg\mathbf{X}_2^{(\mu_c,\nu_c)}$	certain instance set on label l_c
$\hat{\mathbf{y}}_{ic}$	pseudo-label of instance $\mathbf{x}_i$ on label l_c learnt by f_1
$\phi_c(\mathbf{x}_i)$	high-dimensional representation of instance $\mathbf{x}_i$ gives a label-specific feature on label l_c
$\hat{C}_c^+$	class center of the pseudo-positive class on label l_c
$\hat{C}_c^-$	class center of the pseudo-negative class on label l_c
$D(\cdot,\cdot)$	Euclidean distance
r_c^+	radius of the pseudo-positive class on label l_c
r_c^-	radius of the pseudo-negative class on label l_c
$\mathbf{x}_i^c$	label-specific feature of instance $\mathbf{x}_i$ on label l_c
p_c^+	pseudo-positive instance weight on label l_c
p_c^-	pseudo-negative instance weight on label l_c
ρ_c	weighted neighborhood difference of $\mathbf{x}_i$ on label l_c
r	instance neighborhood size measured by Euclidean distance $D(\cdot,\cdot)$
$N_c(\mathbf{x}_i)^+$	pseudo-positive instance count in the neighborhood of $\mathbf{x}_i$ on label l_c
$N_c(\mathbf{x}_i)^-$	pseudo-negative instance count in the neighborhood of $\mathbf{x}_i$ on label l_c
UC_i	the number of times that instance $\mathbf{x}_i$ is regarded as the uncertain instance on all label sets
$\overline{UC_j}$	mean of times that instance $\mathbf{x}_j$ is regarded as the uncertain instance on all label sets

3.2. Problem Formulation

Let $\mathbf{X}_1 = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ denotes a set of multi-label instances with the union of known logical label information $\mathbf{Y}_1 = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n\}$; $\mathbf{X}_2$ denotes the unseen instances. The logical label of $\mathbf{x}_i$ to label set $\{l_1, \ldots, l_c, \ldots, l_m\}$ is denoted as $\mathbf{y}_i = \{y_{i1}, \ldots, y_{ic}, \ldots, y_{im}\}$ where $y_{ic} = 1$ holds (positive class) if $\mathbf{x}_i$ is associated with label l_c, and $y_{ic} = 0$ (negative class) otherwise. The numerical label of instance $\mathbf{x}_i$ is denoted by $\mathbf{u}_i = \{u_{i1}, \ldots, u_{ic}, \ldots, u_{im}\} \in [0,1]^m$, which satisfies $\sum_c u_{ic} = 1$. For an arbitrary label l_c, we denote the degree of the membership function related to l_c and non-membership functions unrelated to l_c as $IFN_c(\mathbf{x}_i) = (\mu_c(\mathbf{x}_i), \nu_c(\mathbf{x}_i))$, respectively, where $\mu_c(\mathbf{x}_i) \in [0,1]$, $\nu_c(\mathbf{x}_i) \in [0,1]$, and $0 \leqslant \mu_c(\mathbf{x}_i) + \nu_c(\mathbf{x}_i) \leqslant 1$. The uncertain instances set on the c-th label is denoted as $\mathbf{X}_2^{(\mu_c, \nu_c)}$. Uncertain instances set on all labels are denoted as $\mathbf{X}_2^{(\mu,\nu)}$, and certain instances set on all labels are complementary of $\mathbf{X}_2^{(\mu,\nu)}$, denoted as $\neg\mathbf{X}_2^{(\mu,\nu)}$ (i.e., $\mathbf{X}_2 = \mathbf{X}_2^{(\mu,\nu)} \cup \neg\mathbf{X}_2^{(\mu,\nu)}$). Our goal was to identify uncertain classifications from logical label-based learning and improve the performance with numerical label-based learning.

3.3. Basic Idea

IFTWLE is an implementation of the TAO model for unseen instances $\mathbf{X}_2$ (see Figure 1). Firstly, the trisecting was realized by a logical label-based model (denoted as $f_1(\cdot)$) with an intuitionistic fuzzy number (denoted as (μ, ν)). An instance selection principle was developed to identify the uncertain instances. Secondly, the acting was realized on all instances with different strategies, with label enhancements on uncertain instances (denoted as $f_2(\cdot)$), adopting the classifications otherwise. Finally, we conducted outcome evaluations on the deduced classifications.

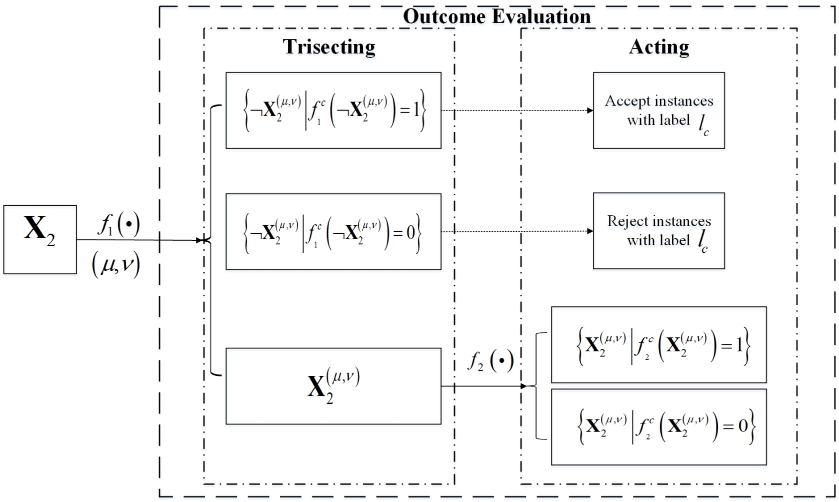

Figure 1. Pipeline of IFTWLE. It follows the trisecting–acting–outcome framework.

For a non-trivial solution, the three-way classification result of the predicted multi-label set (i.e., $\hat{\mathbf{Y}}_2^*$) on $\mathbf{X}_2$ is defined as:

$$\hat{\mathbf{Y}}_2^* = \begin{cases} f_1(\neg\mathbf{X}_2^{(\mu,\nu)}), & \mathbf{x} \in \neg\mathbf{X}_2^{(\mu,\nu)} \\ f_2(\mathbf{X}_2^{(\mu,\nu)}), & \mathbf{x} \in \mathbf{X}_2^{(\mu,\nu)} \end{cases} \tag{8}$$

where $f_2(\mathbf{X}_2^{(\mu,\nu)})$ refers to the predicted multi-label sets of deferred instances with large label uncertainty degree and $f_1(\neg\mathbf{X}_2^{(\mu,\nu)})$ refers to the predicted multi-label sets of certain instances with little label uncertainty degrees.

Figure 2 describes the pipeline on the instance selection principle. By taking a problem transformation view, we assigned the membership function ($\mu_c(\mathbf{x}_i)$) and non-membership function ($\nu_c(\mathbf{x}_i)$) for all unseen instances on an arbitrary label l_c, which was then incorporated to search for the candidate uncertain instances denoted as $\mathbf{X}_2^{(\mu_c,\nu_c)}$. Final uncertain instances ($\mathbf{X}_2^{(\mu,\nu)}$) were aggregated by considering the distributions of candidate uncertain instances across all labels. In what follows, we will elaborate the details from Sections 3.4–3.6.

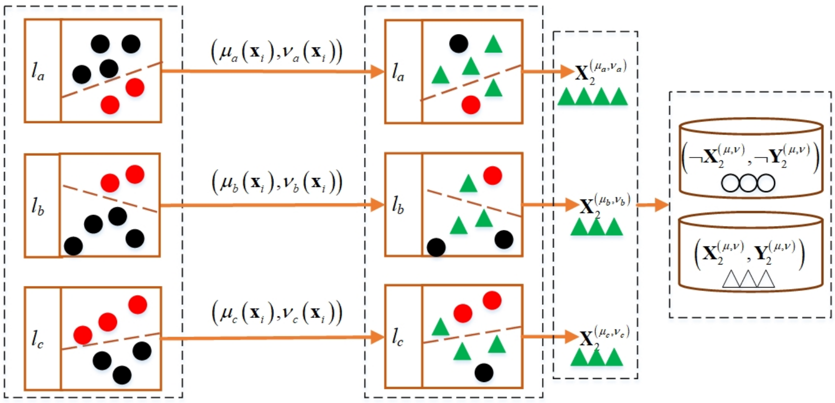

Figure 2. Illustration for the instance selection principle in IFTWLE: we used six instances to explain our processing. After employing $f_1(\cdot)$ generated by label-specific learning, we employed an intuitionistic fuzzy number for each label $((\mu_a(\mathbf{x}_i), \nu_a(\mathbf{x}_i))$ for l_a for example). The processed red and black circles refer to the instances that are classified as positive and negative class with limited uncertainty, and the green triangles are recognized as candidate uncertain instances. The final three uncertain instances (represented by hollow triangles) are denoted as $\mathbf{X}_2^{(\mu,\nu)}$, which conducts label enhancement afterwards.

3.4. Intuitionistic Fuzzy Membership Assignment

Although fuzzy membership assignment is capable of measuring the concept vagueness, it has the following drawbacks for multi-label classification:

- Regardless of the concrete membership function definition, it can only describe the closeness of an instance belonging to a concept. The distribution of the heterogeneous class is thus neglected, which is of great importance in finding uncertain instances.
- Multi-label data have some specialized characteristics, such as an imbalanced class. With the membership function only, the model cannot utilize such information effectively, which leads to the degeneration of model generality.

In [45], the intuitionistic fuzzy set figures out the support vectors from instances and improves the generalization of the support vector machine. In our case, instances have obtained the pseudo-label by conducting label-specific learning (i.e., $f_1(\cdot)$). Therefore we assume a desirable hyperplane is available. This means the possibility of instances with noisy labels is less likely to occur, and misclassified instances are both far from the class center and circled by heterogeneous instances. Therefore, from the perspective of labels, such instances are compatible with the intuitionistic fuzzy set. The degrees of membership and non-membership functions for each unseen instance take the problem transformation view and are determined in a label-specific way. Without losing generality, we consider the construction of $\mu_c(\mathbf{x}_i)$ and $\nu_c(\mathbf{x}_i)$ on label l_c.

3.4.1. Membership Function $\mu_c(\mathbf{x}_i)$

Let $\hat{\mathbf{Y}}_2$ be the pseudo-label set of the unseen instance set $\mathbf{X}_2$ learnt from the LLSF model (f_1); the membership of instances is determined by the relative similarity against the predicted class and other classes. In other words, the membership function of an instance is larger if both the relative distance to the predicted class is smaller and the relative distance to the other classes is larger. Using the class center as a representative, for two instances that are both pseudo-positively associated (i.e., $\hat{y}_{ic} = \hat{y}_{jc} = 1$), our goal can be formally written as:

$$\mu_c(\mathbf{x}_i) > \mu_c(\mathbf{x}_j) \quad if \quad \frac{D\Big(\phi_c(\mathbf{x}_i), \hat{\mathbf{C}}_c^+\Big)}{r_c^+ + \epsilon} - \frac{D\Big(\phi_c(\mathbf{x}_i), \hat{\mathbf{C}}_c^-\Big)}{r_c^- + \epsilon} < \frac{D\Big(\phi_c(\mathbf{x}_j), \hat{\mathbf{C}}_c^+\Big)}{r_c^+ + \epsilon} - \frac{D\Big(\phi_c(\mathbf{x}_j), \hat{\mathbf{C}}_c^-\Big)}{r_c^- + \epsilon} \tag{9}$$

where $D\Big(\phi_c(\mathbf{x}_i), \hat{\mathbf{C}}_c^+\Big)$ and $D\Big(\phi_c(\mathbf{x}_i), \hat{\mathbf{C}}_c^-\Big)$ are the distances of instance $\mathbf{x}_i$ to the pseudo-positive class and pseudo-negative class, respectively. They are defined as:

$$D\Big(\phi_c(\mathbf{x}_i), \hat{\mathbf{C}}_c^+\Big) = \Big\|\phi_c(\mathbf{x}_i) - \hat{\mathbf{C}}_c^+\Big\| \tag{10}$$

$$D\Big(\phi_c(\mathbf{x}_i), \hat{\mathbf{C}}_c^-\Big) = \Big\|\phi_c(\mathbf{x}_i) - \hat{\mathbf{C}}_c^-\Big\| \tag{11}$$

where $\phi_c(\mathbf{x}_i)$ represents the high-dimensional representation of instance $\mathbf{x}_i$ given the label-specific feature w.r.t. l_c, and $\|\cdot\|$ is the distance between the instance and the corresponding pseudo-class center. Suppose $K(\mathbf{x}_i^c, \mathbf{x}_j^c)$ denotes a kernel function on the label-specific feature w.r.t. l_c, the inner product distance is expanded as:

$$\Big\|\phi_c(\mathbf{x}_i) - \phi_c(\mathbf{x}_j)\Big\| = \sqrt{K(\mathbf{x}_i^c, \mathbf{x}_i^c) - 2K(\mathbf{x}_i^c, \mathbf{x}_j^c) + K(\mathbf{x}_j^c, \mathbf{x}_j^c)} \tag{12}$$

$\hat{\mathbf{C}}_c^+$ and $\hat{\mathbf{C}}_c^-$ are the class centers of the pseudo-positive and pseudo-negative classes on label l_c. The average value is measured by all the pseudo-positive instances that predicted the pseudo-label was 1 on label l_c, denoted as $\hat{\mathbf{C}}_c^+$; the average value is measured by all the pseudo-negative instances that predicted pseudo-label was 0 on label l_c, denoted as $\hat{\mathbf{C}}_c^-$.

$$\hat{\mathbf{C}}_c^+ = \frac{1}{l_c^+} \sum_{\hat{y}_{ic}=1} \phi_c(\mathbf{x}_i) \tag{13}$$

$$\hat{\mathbf{C}}_c^- = \frac{1}{l_c^-} \sum_{\hat{y}_{ic}=0} \phi_c(\mathbf{x}_i) \tag{14}$$

where $l_c^+ = |\{\mathbf{x}_i | \hat{y}_{ic} = 1\}|$ and $l_c^- = |\{\mathbf{x}_i | \hat{y}_{ic} = 0\}|$ denote the pseudo-positive instance count and pseudo-negative instance count w.r.t. label l_c, respectively.

r_c^+ and r_c^- are the radii of the pseudo-positive and pseudo-negative class on label l_c, which can be quantified as:

$$r_c^+ = \max_{y_{ic}=1} \Big\|\phi_c(\mathbf{x}_i) - \hat{\mathbf{C}}_c^+\Big\| \tag{15}$$

$$r_c^- = \max_{y_{ic}=0} \Big\|\phi_c(\mathbf{x}_i) - \hat{\mathbf{C}}_c^-\Big\| \tag{16}$$

By substituting Equations (13) and (14) into Equations (15) and (16), based on Equation (12), we have:

$$r_c^+ = \max_{\hat{y}_{ic}=1} \sqrt{K(\mathbf{x}_i^c, \mathbf{x}_i^c) - \frac{2}{l_c^+} \sum_{\hat{y}_{jc}=1} K(\mathbf{x}_i^c, \mathbf{x}_j^c) + \frac{1}{l_c^{+2}} \sum_{\hat{y}_{mc}=1} \sum_{\hat{y}_{nc}=1} K(\mathbf{x}_m^c, \mathbf{x}_n^c)} \tag{17}$$

$$r_c^- = \max_{\hat{y}_{ic}=0} \sqrt{K(\mathbf{x}_i^c, \mathbf{x}_i^c) - \frac{2}{l_c^-} \sum_{\hat{y}_{jc}=0} K(\mathbf{x}_i^c, \mathbf{x}_j^c) + \frac{1}{l_c^{-2}} \sum_{\hat{y}_{mc}=0} \sum_{\hat{y}_{nc}=0} K(\mathbf{x}_m^c, \mathbf{x}_n^c)} \tag{18}$$

one can infer that calculation of $D\Big(\phi_c(\mathbf{x}_i), \hat{\mathbf{C}}_c^-\Big)$ is costly if $\mathbf{x}_i$ is pseudo-positive on l_c. For simplicity, we use $D\Big(\phi_c(\mathbf{x}_i), \hat{\mathbf{C}}_c\Big)$ instead. In other words, we examine the dissimilarity of an instance to the class center of all instances.

For multi-label cases, the positive class is the minority class, whereas the negative class is the majority class. The imbalanced class distribution results in the different contributions to the membership function. The rationality is that the location of the class center of the negative class has a much lower empirical risk than the positive class, and the empirical risk for the center of the positive class becomes higher as the count for instances with the positive class becomes smaller. Therefore, we introduce symbol p_c^+ and p_c^- to represent the pseudo-positive instance weight and pseudo-negative instance weight w.r.t. label l_c. Any l_c, is defined as:

$$p_c^+ = \frac{2 \times l_c^+}{l_c^+ + Card(\mathbf{X}_2)} \tag{19}$$

$$p_c^- = \frac{2 \times l_c^-}{l_c^- + Card(\mathbf{X}_2)} \tag{20}$$

where $Card(\mathbf{X}_2) = |\mathbf{X}_2|$ refers to the instance number in the instance set $\mathbf{X}_2$. For each unseen instance $\mathbf{x}_i$, the degree of membership $\mu_c(\mathbf{x}_i)$ is defined as:

$$\mu_c(\mathbf{x}_i) = \begin{cases} 1 - \left(p_c^+ \times \frac{D(\phi_c(\mathbf{x}_i), \hat{\mathbf{C}}_c^+)}{r_c^+ + \epsilon} + (1 - p_c^+) \times \frac{D(\phi_c(\mathbf{x}_i), \hat{\mathbf{C}}_c)}{r_c + \epsilon} \right) & \hat{y}_{ic} = 1 \\ 1 - \left(p_c^- \times \frac{D(\phi_c(\mathbf{x}_i), \hat{\mathbf{C}}_c^-)}{r_c^- + \epsilon} + (1 - p_c^-) \times \frac{D(\phi_c(\mathbf{x}_i), \hat{\mathbf{C}}_c)}{r_c + \epsilon} \right) & \hat{y}_{ic} = 0 \end{cases} \tag{21}$$

where $\epsilon \to 0^+$ is an adjustable parameter, r_c^+, r_c^-, r_c and $\hat{\mathbf{C}}_c^+$, $\hat{\mathbf{C}}_c^-$, $\hat{\mathbf{C}}_c$ are the radius and class centers of the pseudo-positive class, pseudo-negative class, and all unseen instances on label l_c determined by $f_1^c(\cdot)$. Figure 3 shows an example describing the effect of Equation (21), where both $\mu_c(\mathbf{x}_A) > \mu_c(\mathbf{x}_C)$ and $\mu_c(\mathbf{x}_D) > \mu_c(\mathbf{x}_B)$ hold.

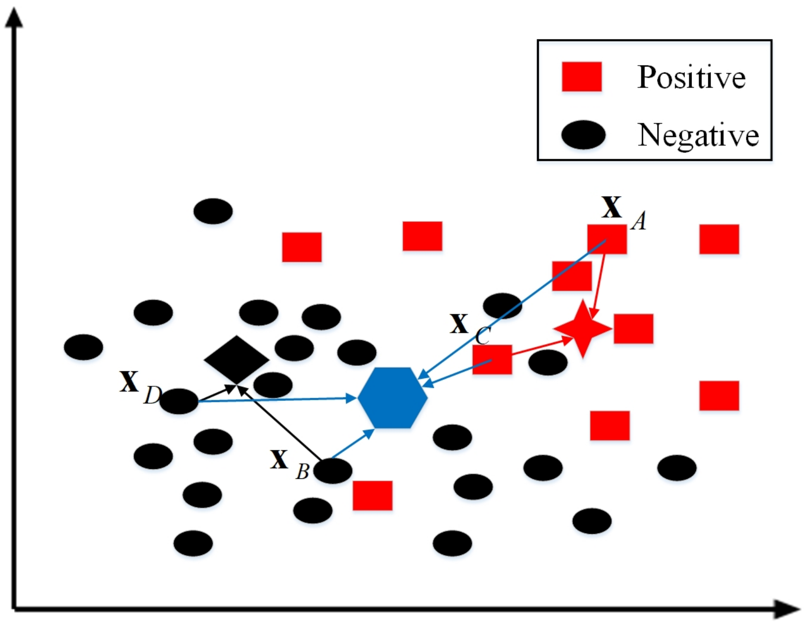

Figure 3. Computing $\mu_c(\mathbf{x}_i)$: The red blocks and black circles represent the pseudo-positive and pseudo-negative w.r.t. label l_c. The red four-angle star and black diamond represent the class center of pseudo-positive and pseudo-negative instances on label l_c. The blue hexagon represents the center for all included instances on label l_c. Instances $\mathbf{x}_A$ and $\mathbf{x}_C$ are two randomly selected instances with the pseudo-positive label on l_c, whereas instances $\mathbf{x}_B$ and $\mathbf{x}_D$ are two randomly selected instances with the pseudo-negative label on l_c. The blue, red, and black lines that connect the instances with the blue hexagon, red four-angle star, and black diamond represent the distances of instances to the instance center, pseudo-positive class center, and pseudo-negative class center, respectively.

3.4.2. Non-Membership Function $\nu_c(\mathbf{x}_i)$

The non-membership of instances is determined by the following two factors. Firstly, it should be impacted by the dissimilarity and similarity of the instance to the predicted class and other classes. It means the non-membership is larger if an instance is located in the region where the probability of being other classes is larger. Secondly, it should be impacted by the heterogeneous instance distribution within the neighborhood. One can infer that the non-membership is larger if an instance is surrounded by instances from other classes. We assume membership is inversely proportional to non-membership, denoted as:

$$\nu_c(\mathbf{x}_i) \propto 1 - \mu_c(\mathbf{x}_i) \tag{22}$$

The imbalanced class distribution [46] implies that the contributions of heterogeneous instances are label-dependent. In other words, for two instances, $\mathbf{x}_i$, and $\mathbf{x}_j$, with the same number of heterogeneous instances in their corresponding neighbours, the non-membership of $\mathbf{x}_i$ is larger than $\mathbf{x}_j$ if $\mathbf{x}_i$ belongs to the pseudo-negative class whereas $\mathbf{x}_j$ belongs to the pseudo-positive class. To implement our assumption, we introduced two symbols, n_c^+ and n_c^-, to represent the prior probability of being pseudo-positive and pseudo-negative, respectively.

$$n_c^+ = \frac{l_c^+}{Card(\mathbf{X}_2)} \tag{23}$$

$$n_c^- = \frac{l_c^-}{Card(\mathbf{X}_2)} \tag{24}$$

The two prior probabilities are incorporated to quantify the contribution of the heterogeneous instance within the neighborhood. From the perspective of the neighborhood, the non-membership degree is larger if more heterogeneous instances are included, and smaller if more homogeneous instances are included. Therefore, the weighted neighborhood difference of $\mathbf{x}_i$ on label l_c (i.e., $(\rho_c(\mathbf{x}_i))$) is defined as:

$$\rho_c(\mathbf{x}_i) = \begin{cases} \frac{n_c^+ N_c(\mathbf{x}_i)^-}{n_c^+ N_c(\mathbf{x}_i)^- + n_c^- N_c(\mathbf{x}_i)^+} & \hat{y}_{ic} = 1 \\ \frac{n_c^- N_c(\mathbf{x}_i)^+}{n_c^+ N_c(\mathbf{x}_i)^- + n_c^- N_c(\mathbf{x}_i)^+} & \hat{y}_{ic} = 0 \end{cases} \tag{25}$$

where $N_c(\mathbf{x}_i)^+ = \left|\{\mathbf{x}_j | \mathbf{x}_j \in N_c(\mathbf{x}_i) \wedge \hat{y}_{jc} = 1\}\right|$ denotes the pseudo-positive instance count in the r-neighborhood of $\mathbf{x}_i$ on label l_c ($r > 0$). $r > 0$ is an adjustable parameter and $N_c(\mathbf{x}_i)^- = \left|\{\mathbf{x}_j | \mathbf{x}_j \in N_c(\mathbf{x}_i) \wedge \hat{y}_{jc} = 0\}\right|$ denotes the pseudo-negative instance count in the γ-neighborhood of $\mathbf{x}_i$ on label l_c. We assume the non-membership degree is proportional to the weighted neighborhood difference, denoted as:

$$\nu(\mathbf{x}_i) \propto \rho(\mathbf{x}_i) \tag{26}$$

Based on assumptions (22) and (26), we define the non-membership degree as:

$$\nu_c(\mathbf{x}_i) = (1 - \mu_c(\mathbf{x}_i))\rho_c(\mathbf{x}_i) \tag{27}$$

It is easy to validate $0 \leqslant \mu_c(\mathbf{x}_i) + \nu_c(\mathbf{x}_i) \leqslant 1$ holds. Here, we offer some explanations. The $\mu_c(\mathbf{x}_i)$ receives the largest value when it is the center of the plausible pseudo-class, and it reaches 0 only when it is the farthest instance to the class center of all instances simultaneously. The $\nu_c(\mathbf{x}_i)$ itself is smaller than 1, as the two components are all less than 1. As $\rho_c(\mathbf{x}_i)$ is smaller than 1, the $\nu_c(\mathbf{x}_i)$ is no larger than $1 - \mu_c(\mathbf{x}_i)$, which means $0 \leqslant \mu_c(\mathbf{x}_i) + \nu_c(\mathbf{x}_i) \leqslant 1$ holds.

By referring to Equation (6), we define the hesitation degree as:

$$\pi_c(\mathbf{x}_i) = 1 - \mu_c(\mathbf{x}_i) - \nu_c(\mathbf{x}_i) \tag{28}$$

3.5. Label-Level Instance Selection

Having the membership/non-membership degrees, the unseen instances ($\mathbf{X}_2$) with the intuitionistic fuzzy membership ($\mathbf{IFX}_2$) are denoted as:

$$\mathbf{IFX}_2 = \{(\mathbf{x}_1, \hat{\mathbf{y}}_1, \mu_1, \nu_1), (\mathbf{x}_2, \hat{\mathbf{y}}_2, \mu_2, \nu_2), \ldots, (\mathbf{x}_k, \hat{\mathbf{y}}_k, \mu_k, \nu_k)\}$$

where $\mu_i = (\mu_1(\mathbf{x}_i), \mu_2(\mathbf{x}_i), \ldots, \mu_m(\mathbf{x}_i))$ and $\nu_i = (\nu_1(\mathbf{x}_i), \nu_2(\mathbf{x}_i), \ldots, \nu_m(\mathbf{x}_i))$ denote the degrees of membership and non-membership functions w.r.t. $\mathbf{x}_i$ across all labels, respectively.

In terms of each label, we can select the instances with uncertain classifications (denoted as $\mathbf{X}_2^{(\mu_c,\nu_c)}$) as:

$$\mathbf{X}_2^{(\mu_c,\nu_c)} = \{\mathbf{x}_i | \mu_c(\mathbf{x}_i) > \nu_c(\mathbf{x}_i) \wedge \nu_c(\mathbf{x}_i) > 0\} \quad (29)$$

$\mathbf{X}_2^{(\mu_c,\nu_c)}$ can be interpreted as candidate uncertain instances. For example, assume four unseen samples with pseudo-labels, i.e., $\mathbf{x}_A$, $\mathbf{x}_B$, $\mathbf{x}_C$ and $\mathbf{x}_D$ in Figure 4. It is more likely that $\mathbf{x}_B$ and $\mathbf{x}_C$ belong to $\mathbf{X}_2^{(\mu_c,\nu_c)}$ than $\mathbf{x}_A$ and $\mathbf{x}_D$, and they are more close to the hyperplane. As instances with small separation margins tend to be more uncertain, label enhancement on $\mathbf{x}_C$ and $\mathbf{x}_D$ will be more informative than on $\mathbf{x}_A$ and $\mathbf{x}_B$, given the pseudo-label distribution on label l_c. One should be aware that $\mathbf{x}_C$ is less likely to be considered if we only apply the traditional fuzzy set model, as the affiliation degree to the positive class is much larger than that of the negative class. However, it will be conducive to enhancing generality if $\mathbf{x}_C$ is selected, as there are two instances with negative classes in its neighbours. Hence, we can select instance $\mathbf{x}_C$ via the intuitionistic fuzzy set.

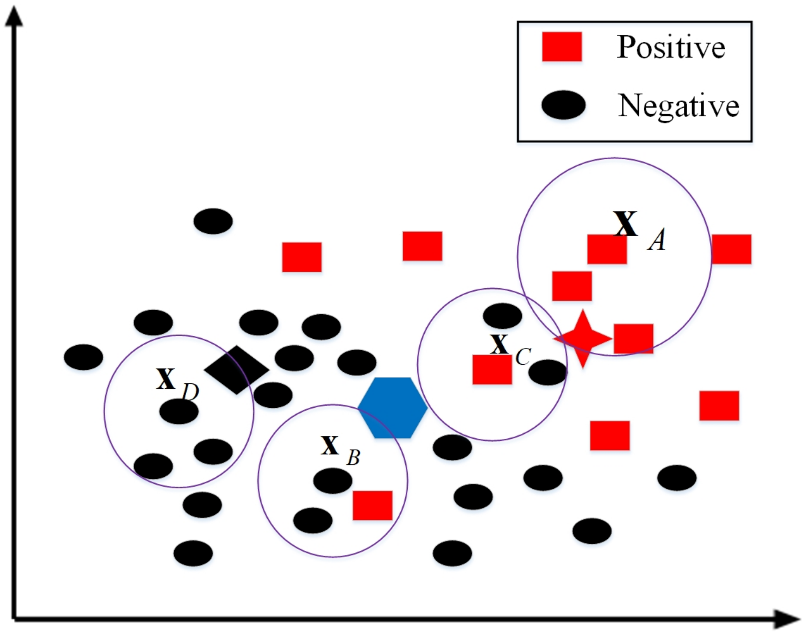

Figure 4. Selection of uncertain candidates on label l_c: The red blocks and black circles represent the instances with pseudo-positive and pseudo-negative on label l_c. The red four-angle star and black diamond represent the class center of pseudo-positive and pseudo-negative instances on label l_c, respectively. The blue hexagon represents the center for all included instances on label l_c. The purple circle represents the region of the neighborhood. Based on (29), $\mathbf{x}_B$ and $\mathbf{x}_C$ will be selected (i.e., $\mathbf{x}_B, \mathbf{x}_C \in \mathbf{X}_2^{(\mu_c,\nu_c)}$).

3.6. Instance-Level Instance Selection

Now we determine which instances should be selected for the label enhancement. Since label enhancement works at the instance level, a straightforward notion is to select the instances that are uncertain across most labels. In this paper, we introduce the symbol UC_i to represent the number of accumulated times that satisfy $\mathbf{x}_i \in \mathbf{X}_2^{(\mu_c,\nu_c)}$ for any l_c. It is defined as:

$$UC_i = \sum_c \left[\left[\mathbf{x}_i \in \mathbf{X}_2^{(\mu_c,\nu_c)}\right]\right] \quad (30)$$

where the notation $[\![\cdot]\!]$ equals 1 if the condition meets and 0 otherwise. Based on the UC, the instances to be enhanced ($\mathbf{X}_2^{(\mu,\nu)}$) can be computed as:

$$\mathbf{X}_2^{(\mu,\nu)} = \left\{\mathbf{x}_i \middle| UC_i \geqslant \overline{UC_j} \wedge \mathbf{x}_j \in \mathbf{X}_2 \wedge UC_j > 0\right\} \tag{31}$$

where the notation $\overline{UC_j}$ refers to the average non-zero UC_j. This means that an instance will be enhanced only when it is recognized as uncertain in most labels.

3.7. Complexity Analysis

We summarize the procedures of IFTWLE in Algorithm 1. The most time-consuming step is the calculation of pair-wise distances within all unseen instances (i.e., r_c). Let k be the average length of label-specific features, n be the unseen instance count, and m be the label count, then the calculation for r_c requires $O(n^2mk)$. Therefore, the computational complexity of IFTWLE is much lower than that of training $f_1(\cdot)$ [9] and $f_2(\cdot)$ [16]. However, as the complexity is proportional to the quadratic of the instance count, it is inappropriate to employ this algorithm for large-scale multi-labels. The local estimation of the inner product should be considered to accelerate the instance selection.

Algorithm 1: IFTWLE

Input: instances $\mathbf{X} = \mathbf{X}_1 \cup \mathbf{X}_2$, logical labels $\mathbf{Y}_1$, parameters α, β, and γ, neighborhood size r.
Output: unseen labels $\hat{\mathbf{Y}}_2^*$.

Construct $f_1(\cdot)$ as described in (1).
Generate pseudo-labels $\hat{\mathbf{Y}}_2$ given $\mathbf{X}_2$ as described in (2).
for $c = 1$ *to* m **do**
  Compute r_c^+ and r_c^- as described in (17) and (18).
  for $i = 1$ *to* n **do**
    Compute $\mu_c(\mathbf{x}_i)$ as described in (21).
    Compute $\rho_c(\mathbf{x}_i)$ as described in (25).
    Compute $\nu_c(\mathbf{x}_i)$ as described in (27).
  end
  Compute $\mathbf{X}_2^{(\mu_c,\nu_c)}$ as described in (29).
end
Compute $\mathbf{X}_2^{(\mu,\nu)}$ as described in (31).
$\neg\mathbf{X}_2^{(\mu,\nu)} = \mathbf{X}_2 - \mathbf{X}_2^{(\mu,\nu)}$.
if $\mathbf{X}_2^{(\mu,\nu)} \neq \varnothing$ **then**
  Construct $f_2(\cdot)$ as described in (3).
  Generate pseudo-labels $\hat{\mathbf{Y}}_2^{(\mu,\nu)}$ as described in (4).
end
Generate $\hat{\mathbf{Y}}_2^*$ as described in (8).

4. Experiments

4.1. Dataset Characteristics

To demonstrate the effectiveness and efficiency of the proposed model, we compared the classification performances on eight multi-label benchmarks from Mulan (http://mulan.sourceforge.net/datasets.html) (accessed on 1 March 2022) and Meka (http://waikato.github.io/meka/datasets/) (accessed on 1 March 2022) on domains including audio, music, text, biology, and images. The selected benchmarks were either small or moderate and were intensively referenced due to their limited baseline performances. In Table 2, for each dataset, "# Instances" means the number of 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, where notation "#" means the number of.

Table 2. Characteristics of data sets.

Data Set	# Instances	# Features	# Labels	# Cardinality	Domain
birds	645	260	19	1.014	audio
emotions	593	72	6	1.869	music
enron	1702	1001	53	3.378	text
genbase	662	1185	27	1.252	biology
languagelog	1460	1004	75	1.18	text
medical	978	1449	45	1.245	text
scene	2407	294	6	1.074	image
yeast	2417	103	14	4.237	biology

4.2. Evaluation Metrics

We adopted six evaluation metrics (*Hamming Loss*, *Ranking Loss*, *One Error*, *Coverage*, *Average Precision*, and *Micro F1*) [47] to evaluate the classification performance. Except for the last two, which achieve better performance when the values are larger, the remaining metrics obtain better performances when values are smaller. Let $\mathbf{Y_i}$ and $\neg\mathbf{Y_i}$ denote the relevant and irrelevant label sets in ground-truth, and n be the unseen instances count, then the formulas of metrics are enumerated as:

(1) *Hamming Loss* (abbreviated as *Hl*) evaluates the average difference between predictions and ground-truth (see Formula (32)). The smaller the value of the *Hamming Loss*, the better the performance of the algorithm.

$$Hl = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{l}Card(f(\mathbf{x}_i)\Delta\mathbf{Y}_i) \tag{32}$$

where Δ is the set symmetric difference and $Card(\cdot)$ is the set cardinality.

(2) *Ranking Loss* (abbreviated as *Rkl*), evaluates the fraction that the irrelevant label ranks before the relevant label in the label predictions (see Formula (33)). The smaller the value of the *Ranking Loss*, the better the performance of the algorithm.

$$Rkl = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{Card(\mathbf{Y}_i)Card(\neg\mathbf{Y}_i)}\times Card(\{(l_a,l_b)|rank_i(l_a) > rank_i(l_b)\wedge(l_a,l_b)\in\mathbf{Y}_i\times\neg\mathbf{Y}_i\}) \tag{33}$$

where $rank_i(l_j)$ denotes the ranking position in ascending order for the j-th label on the i-th instance. $Card(\cdot)$ is the set cardinality.

(3) *One Error* (abbreviated as *Oe*) evaluates the average fraction that the label ranking—first in prediction—is actually the irrelevant label (see Formula (34)). The smaller the value of *One Error*, the better the performance of the algorithm.

$$Oe = \frac{1}{n}\sum_{i=1}^{n}\left[\!\!\left[\left(\arg\min_{l_j} rank_i(l_j)\right)\notin\mathbf{Y}_i\right]\!\!\right] \tag{34}$$

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

(4) *Coverage* (abbreviated as *Cvg*) evaluates the average fraction for inclusion of all ground-truth labels in the ranking of label predictions (see Formula (35)). The smaller the value of *Coverage*, the better the performance of the algorithm.

$$Cvg = \frac{1}{n}\sum_{i=1}^{n}\max_{l_j\in\mathbf{Y}_i} rank_i(l_j) - 1 \tag{35}$$

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

(5) *Average Precision* (abbreviated as *Ap*) evaluates the average precision of the actual relevant label rankings before relevant labels are examined by the label predictions (see Formula (36)). The larger the value of *Average Precision*, the better the performance of the algorithm.

$$Ap = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{Card(\mathbf{Y}_i)}\sum_{l_j\in\mathbf{Y}_i}\frac{Card(\{l_s\in\mathbf{Y}_i|rank_i(l_s)\leqslant rank_i(l_j)\})}{rank_i(l_j)} \tag{36}$$

where $Card(\cdot)$ is the set cardinality.

4.3. Experimental Settings

We examined whether IFTWLE gained superiority over state-of-the-art algorithms learnt on logical labels only. For this reason, we compared IFTWLE with LLSF, multi-label k-nearest neighbour (MLkNN), multi-label learning with label-specific features (LIFT), MLTSVM, multi-label learning with global and local label correlation (Glocal), and active k-label set ensembles (ACkEL). Detailed settings are as follows.

- LLSF (code available at https://jiunhwang.github.io/ (accessed on 1 March 2022)) [9]: This method learns label-specific feature representations for all labels based on logical labels. It shares an identical structure with $f_1(\cdot)$. Comparing this method, we could examine the contributions of label enhancement. Parameters δ, η are tuned in $\{2^{-10}, 2^{-9}, \ldots, 2^{9}, 2^{10}\}$. The calibrated threshold τ_1 is fixed as 0.5.
- MLkNN (code available at http://www.lamda.nju.edu.cn/code_MLkNN.ashx (accessed on 1 March 2022)) [48]: It learns a conditional probability distribution on all features within the adapted k-neighborhood. The introduction of the neighborhood has some similarities with the components in the non-membership function $\nu(\mathbf{x}_i)$. However, we took one further step and enhanced the results of the uncertain instances. The value k took the empirical value 10.
- LIFT (code available at http://cse.seu.edu.cn/PersonalPage/zhangml/index.htm (accessed on 1 March 2022)) [40]: It learned different feature representations to determine the label association. It was the first trial of label-specific learning for multi-label classification. By comparing with this method, we could verify whether label enhancement improves label-specific learning. The ratio parameter is searched in $\{0.1, 0.2, \ldots, 0.5\}$.
- MLTSVM (code available at http://www.optimal-group.org/Resource/MLTSVM.html (accessed on 1 March 2022)) [11]: It learns distance difference based on multiple nonparallel hyperplanes. The twin support vector machine is a variant of the support vector machine; we trained the enhanced model by a linear support vector machine. The penalty and kernel parameter were searched in $\{2^{-6}, 2^{-5}, \ldots, 2^{5}, 2^{6}\}$ and $\{2^{-4}, 2^{-3}, \ldots, 2^{3}, 2^{4}\}$, respectively.
- Glocal (code available at http://www.lamda.nju.edu.cn/code_Glocal.ashx (accessed on 1 March 2022)) [49]: It learns a mapping from the feature space to latent labels via a low-rank decomposition. It is the initial attempt to simultaneously leverage both global and local label correlations. The similarity is that we also consider global label correlation in a pairwise fashion. By comparing with this work, we can examine whether label importance is superior to local label correlation. The penalty λ takes the empirical value 1.
- ACkEL (code available at https://github.com/xuwangfmc/AkEL (accessed on 1 March 2022)) [50]: It takes an ensemble strategy on the k-label set optimized by class separability and class uncertainty simultaneously. It is a revised version of the classic k-label set algorithm, which is assumed to gain a robust performance. By comparing with this work, we can examine whether the strategy, combining the pairwise label correlation and label importance, could gain superiority over a high-order label correlation. A one-versus-all multi-class strategy was conducted on each label set;

parameters σ, β were searched in $\{2^{-3}, 2^{-2}, \ldots, 2^{10}, 2^{11}\}$ and $\{0.1, 0.3, 0.5, 0.7, 0.9\}$, respectively. The size for each label set was fixed as 3.

- Proposed method. There are three groups of parameters. The parameters in constructing $f_1(\cdot)$ and $f_2(\cdot)$ take the recommended settings, as declared in [9] and [16]. For an intuitionistic fuzzy membership assignment, all components were automatically determined by data characteristics except for neighborhood radius r. It was searched in $\{0.01, 0.1\}$ via five-fold cross-validation.

We considered five evaluations [47] including *Hamming Loss*, *One Error*, *Coverage*, *Ranking Loss*, and *Average Precision*. Except for *Average Precision*, which obtains better performance if the metric becomes larger, the others achieve better performances if the metrics become smaller. The experiments were implemented using Matlab R2017b on a desktop PC with an Intel(R) Core i7 processor (2.60 GHz) and 8 GB of RAM. All parameters were selected via five-fold cross-validation.

4.4. Results

We evaluated the classification performances, considering algorithms on five evaluation metrics; we report them in Figures 5–9. The down arrow ↓ means the smaller the metrics is, the better the algorithm performance becomes. In contrast, the up arrow ↑ means the larger the metrics is, the better the algorithm performance becomes.

From the metric view, IFTWLE ranks in first place in 60% of cases ($\frac{3}{5}$) and in second place in 40% of cases. From the dataset view, IFTWLE ranks in first place in 42.5% of cases ($\frac{17}{40}$) and in second place in 17.5% of cases ($\frac{7}{40}$). It receives the best performance on the *Coverage* metric (with first place in 87.5% cases) and worst on the *One Error* metric (with second place in 75% of cases).

The Friedman test [51] was employed to calculate the relative performances among multiple algorithms over selected datasets. Given k comparing algorithms and N datasets, let $Rank_j = (1/N)\sum_{i=1}^{N} r_i^j$ denote the average rank for the j-th algorithm. With the null hypothesis (i.e., H_0) of all algorithms obtaining identical performance, the Friedman statistic F_F is distributed according to the F-distribution with the $k-1$ degree of freedom as the numerator and $(k-1)(N-1)$ degree of freedom as the denominator, denoted as:

$$F_F = \frac{(N-1)\chi_F^2}{N(k-1) - \chi_F^2} \tag{37}$$

where

$$\chi_F^2 = \frac{12N}{k(k+1)}\left[\sum_j Rank_j^2 - \frac{k(k+1)^2}{4}\right] \tag{38}$$

Table 3 presents the Friedman statistics F_F and the corresponding critical value for all evaluation metrics in this setting. The results clearly support that, at the significance level, $\alpha = 0.05$, the null hypothesis (i.e., H_0) of the statistically indistinguishable performance of all algorithms on the considered metrics is rejected. It implies that it is feasible to examine if IFTWLE gains statistical superiority over other comparing algorithms by conducting a post hoc test, such as the Holm [51].

Furthermore, by regarding IFTWLE as the control algorithm, we employed the Holm procedure [51] to explore whether IFTWLE gains significant a performance difference against each of the considered algorithms. Without losing generality, we nominate A_1 as IFTWLE. For the other $k-1$ comparing algorithms (i.e., $A_j (2 \leqslant j \leqslant k)$), we stipulate A_j as the one that has the $j-1$-th largest average ranking over all datasets on a specific evaluation metric. Consequently, we have the test statistic for comparing A_1 (i.e., IFTWLE) with A_j as:

$$z_j = \left(Rank_1 - Rank_j\right) \Big/ \sqrt{\frac{k(k+1)}{6N}} \quad (2 \le j \le k) \tag{39}$$

Let p_j denote the p-value of z_j under a normal distribution. Given the significance level $\alpha = 0.05$, the Holm procedure works in a stepwise manner by checking whether the statistics p_j are smaller than $\alpha/(k-j+1)$ in ascending order of j. Specifically, the Holm procedure continues until there exists a j^*-th step, and j^* denotes the first j, such that $p_j \geqslant \alpha/(k-j+1)$ holds (If $p_j < \alpha/(k-j+1)$ holds for all j, j^* takes the value of $k+1$).

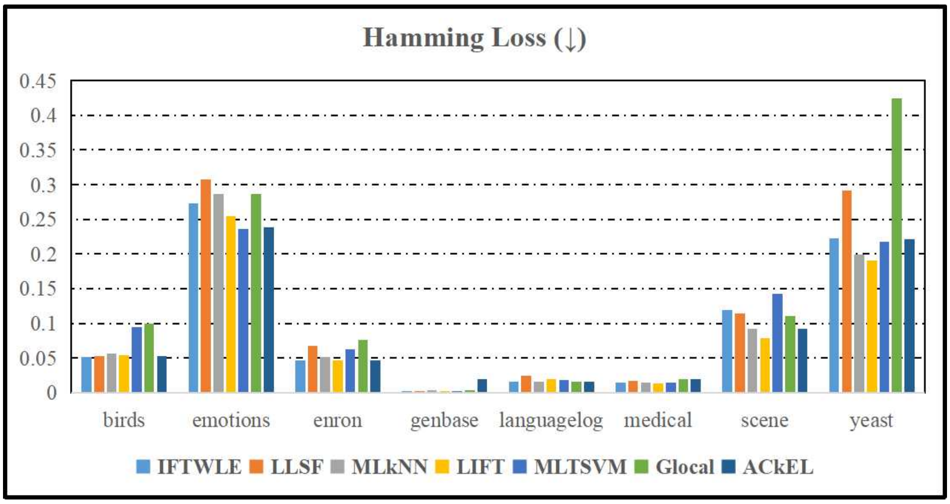

Figure 5. Comparison of each algorithm on the *Hamming Loss* metric. The average rankings of algorithms on this metric: LIFT (2.625) > IFTWLE (3.4375) > AC*k*EL (3.6250) > MLTSVM (3.9375) > ML*k*NN (3.8125) > LLSF (5.0000) > Glocal (5.5625).

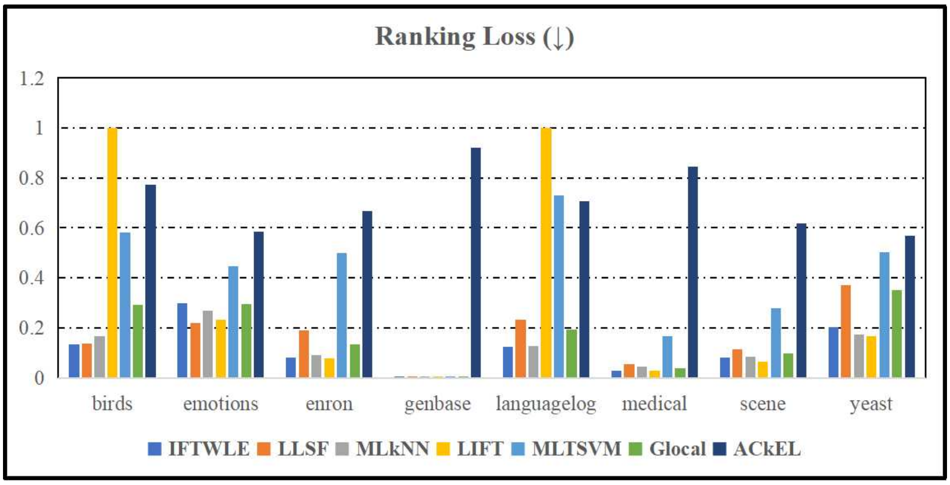

Figure 6. Comparison of each algorithm on the *Ranking Loss* metric. The average rankings of algorithms on this metric: IFTWLE (2.0000) > ML*k*NN (3.1250) = LIFT (3.1250) > Glocal (3.5625) > LLSF (3.6875) > MLTSVM (5.8750) > AC*k*EL (6.6250).

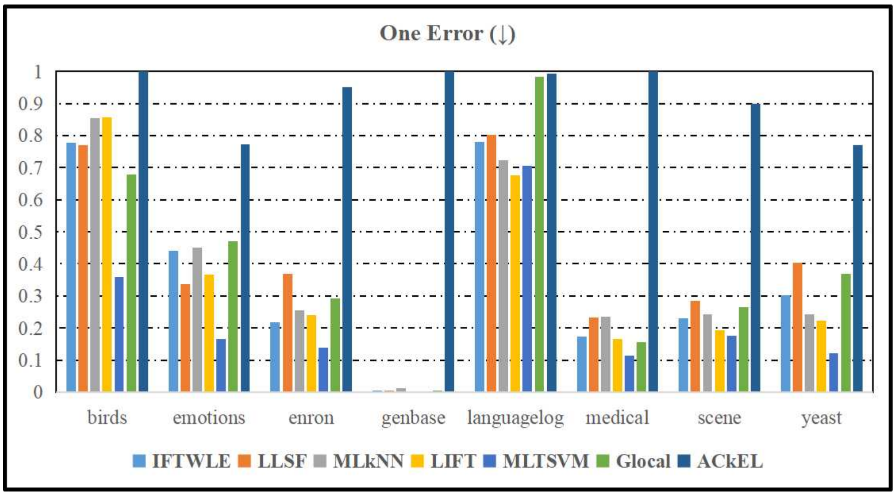

Figure 7. Comparison of each algorithm on the *One Error* metric. The average rankings of algorithms on this metric: MLTSVM (1.2500) > LIFT (2.6250) > IFTWLE (3.6250) > Glocal (4.3750)>ML*k*NN (4.5000) > LLSF (4.6250) > AC*k*EL (7.0000).

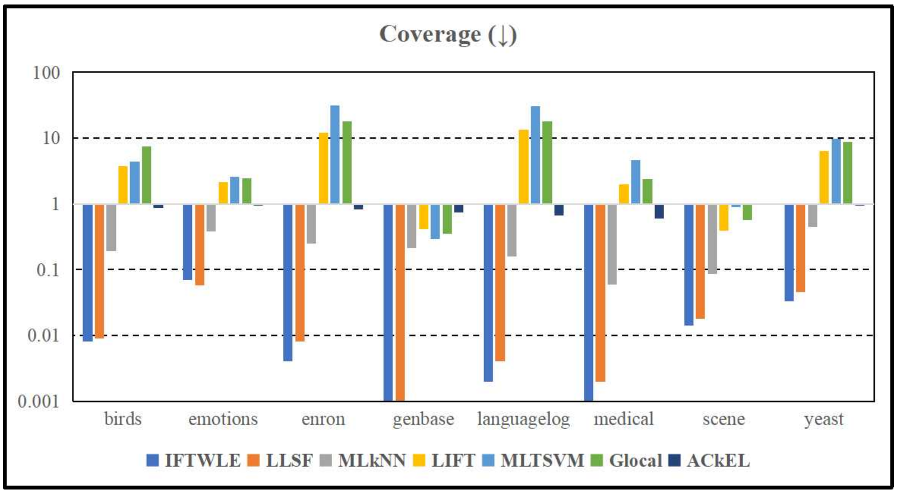

Figure 8. Comparison of each algorithm on the *Coverage* metric. The average rankings of algorithms on this metric: IFTWLE (1.1875) > LLSF (1.8125) > ML*k*NN (3.0000) > AC*k*EL (4.7500) > LIFT (5.0000) > Glocal (5.8750) > MLTSVM (6.3750).

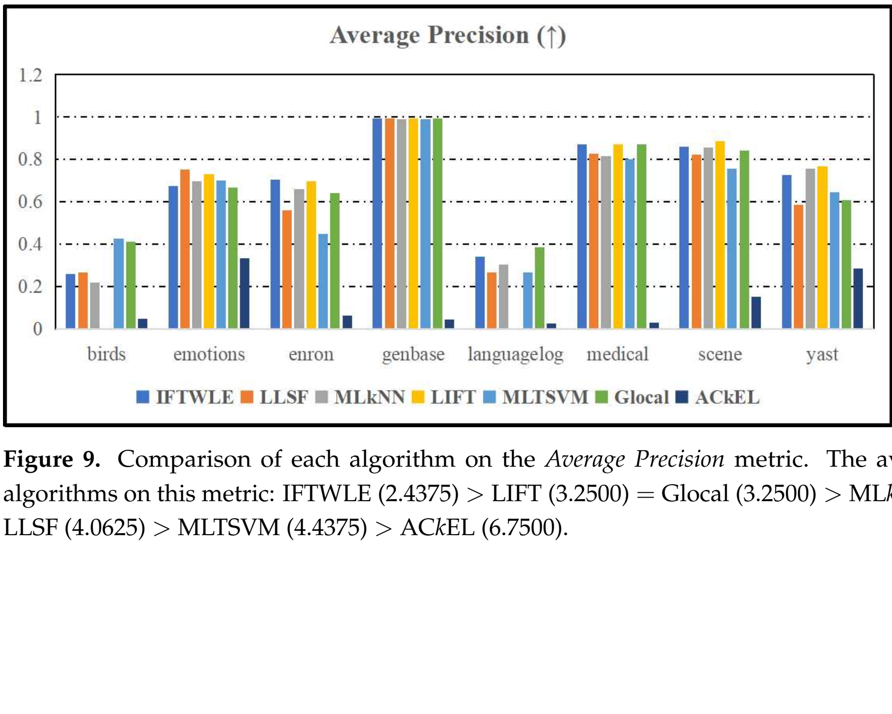

Figure 9. Comparison of each algorithm on the *Average Precision* metric. The average rank of algorithms on this metric: IFTWLE (2.4375) > LIFT (3.2500) = Glocal (3.2500) > ML*k*NN (3.8125) > LLSF (4.0625) > MLTSVM (4.4375) > AC*k*EL (6.7500).

Table 3. Summary of the Friedman statistics F_F ($k = 7$, $N = 8$) and critical values in terms of each evaluation measure (k:# comparing algorithms; N:# data sets).

Metric	F_F	Critical Value ($\alpha = 0.05$)
Hamming Loss	9.991071	
Ranking Loss	27.816964	
One Error	33.214286	2.3239
Coverage	41.852679	
Average Precision	19.473214	

It can be seen in Tables 4–8 that IFTWLE statistically outperforms AC*k*EL on the *Ranking Loss*, *Coverage*, and *Average Precision* metrics, and statistically outperform MLTSVM on the *Ranking Loss* and *Coverage* metrics. IFTWLE achieves the most dominance on Coverage, which is statistically superior over all algorithms except ML*k*NN and LLSF. By finding instances with the most uncertain labels, it is more likely to revise a large proportion of misclassified labels, gaining more improvements in the *Coverage* metric. However, this strategy does not discriminate concerning the relative importance of labels in different instances, which lead to limited improvements in the *One Error* and *Average Precision* metrics.

Table 4. Comparison of IFTWLE (control algorithm) against other comparing algorithms (with the Holm procedure as the post hoc test) at the significant level $\alpha = 0.05$ on the *Hamming Loss metric.*

j	Algorithm	z_j	p	Holm
2	Glocal	−1.9674	0.0491	0.008
3	LLSF	−1.4466	0.1480	0.010
4	MLTSVM	−0.4629	0.6434	0.013
5	ML*k*NN	−0.3472	0.7284	0.017
6	AC*k*EL	−0.1736	0.8622	0.025
7	LIFT	0.7522	1.0000	0.050

Table 5. Comparison of IFTWLE (control algorithm) against other comparing algorithms (with the Holm procedure as the post hoc test) at the significance level $\alpha = 0.05$ on the *Ranking Loss metric.* Algorithms that are statistically inferior to IFTWLE are in bold.

j	Algorithm	z_j	p	Holm
2	**AC*k*EL**	−4.281918	0.000019	0.008
3	**MLTSVM**	−3.587553	0.000334	0.010
4	LLSF	−1.562321	0.118212	0.013
5	Glocal	−1.446594	0.148011	0.017
6	ML*k*NN	−1.041548	0.297621	0.025
7	LIFT	−1.041548	0.297621	0.050

Table 6. Comparison of IFTWLE (control algorithm) against other comparing algorithms (with the Holm procedure as the post hoc test) at the significance level $\alpha = 0.05$ on the *One Error metric.* Algorithms that are statistically inferior to IFTWLE are in bold.

j	Algorithm	z_j	p	Holm
2	**AC*k*EL**	−3.1246	0.0018	0.008
3	LLSF	−0.9258	0.3545	0.010
4	ML*k*NN	−0.8101	0.4179	0.013
5	Glocal	−0.6944	0.4874	0.017
6	LIFT	0.9258	1.0000	0.025
7	MLTSVM	2.1988	1.0000	0.050

Table 7. Comparison of IFTWLE (control algorithm) against other comparing algorithms (with the Holm procedure as the post hoc test) at the significance level $\alpha = 0.05$ on the *Coverage metric*. Algorithms that are statistically inferior to IFTWLE are in bold.

j	Algorithm	z_j	p	Holm
2	**MLTSVM**	−4.802692	0.000002	0.008
3	**Glocal**	−4.339782	0.000014	0.010
4	**LIFT**	−3.529689	0.000416	0.013
5	**AC*k*EL**	−3.2982345	0.000973	0.017
6	ML*k*NN	−1.678049	0.093338	0.025
7	LLSF	−0.578638	0.562834	0.050

Table 8. Comparison of IFTWLE (control algorithm) against other comparing algorithms (with the Holm procedure as the post hoc test) at the significance level $\alpha = 0.05$ on the *Average Precision* metric. Algorithms that are statistically inferior to IFTWLE are in bold.

j	Algorithm	z_j	p	Holm
2	**AC*k*EL**	−3.992599	0.000134	0.008
3	MLTSVM	−1.85164	0.064078	0.010
4	LLSF	−1.504458	0.132464	0.013
5	ML*k*NN	−1.273003	0.203017	0.017
6	LIFT	−0.752229	0.451913	0.025
7	Glocal	−0.752229	0.451913	0.050

5. Conclusions

This paper proposes a novel model called IFTWLE to deal with multi-label classification. Unlike conventional multi-label learning algorithms, which learn models on either logical labels or numerical labels, we integrated the two forms of labels under the three-way decisions umbrella by exploring classification uncertainty. The intuitionistic fuzzy set provides an insightful view into quantifying the label level uncertainty and it determines the uncertain instances in a group decision-making fashion. Comparisons on benchmarks demonstrate that IFTWLE significantly improves classification performance.

In the future, we will examine more combinations of label-specific algorithms and label enhancement algorithms to see whether some guidelines exist. Meanwhile, we will develop advanced instance selection principles by resorting to the optimization theory.

Author Contributions: Conceptualization, formal analysis, writing—original draft preparation, T.Z.; methodology, software, validation, writing—review and editing, Y.Z.; resources, supervision, funding acquisition, D.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research is supported by the National Natural Science Foundation of China grant number 61976158, 61976160, 62076182, 62163016, 62006172, 61906137, and is also partially supported by the Jiangxi ”Double Thousand Plan”, grant number 20212ACB202001.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Zhang, M.L.; Zhou, Z.H. A review on multi-label learning algorithms. *IEEE Trans. Knowl. Data Eng.* **2014**, *26*, 1819–1837. [CrossRef]
2. Gibaja, E.; Ventura, S. A tutorial on multilabel learning. *ACM Comput. Surv.* **2015**, *47*, 1–38. [CrossRef]
3. Liu, W.W.; Shen, X.B.; Wang, H.B.; Tsang, I.W. The emerging trends of multi-label learning. *IEEE Trans. Pattern Anal. Mach. Intell.* 2021, *in press*. [CrossRef] [PubMed]
4. Tabatabaei, S.M.; Dick, S.; Xu, W.S. Toward non-intrusive load monitoring via multi-label classification. *IEEE Trans. Smart Grid.* **2017**, *8*, 26–40. [CrossRef]

5. Fu, H.Z.; Cheng, J.; Xu, Y.W.; Wong, D.W.K.; Liu, J.; Cao, X.C. Joint optic disc and cup segmentation based on multi-label deep network and polar transformation. *IEEE Trans. Med. Imag.* **2018**, *37*, 1597–1605. [CrossRef] [PubMed]
6. Wei, Y.C.; Xia, W.; Lin, M.; Huang, J.S.; Ni, B.B.; Dong, J.; Zhao, Y.; Yan, S.C. HCP: A flexible cnn framework for multi-label image classification. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 1901–1907. [CrossRef]
7. Boutell, M.R.; Luo, J.; Shen, X.; Brown, C.M. Learning multi-label scene classification. *Pattern Recog.* **2004**, *37*, 1757–1771. [CrossRef]
8. Tsoumakas, G.; Vlahavas, I. Random k-labelsets: An ensemble method for multilabel classification. *Lect. Notes Artif. Intell.* **2007**, *4701*, 406–417.
9. Huang, J.; Li, G.R.; Huang, Q.M.; Wu, X.D. Learning label-specific features and class-dependent labels for multi-label classification. *IEEE Trans. Knowl. Data Eng.* **2016**, *28*, 3309–3323. [CrossRef]
10. Wu, Q.Y.; Tan, M.K.; Song, H.J.; Chen, J.; Ng, M.K. ML-FOREST: A multi-label tree ensemble method for multi-label classification. *IEEE Trans. Knowl. Data Eng.* **2016**, *28*, 2665–2680. [CrossRef]
11. Chen, W.J.; Shao, Y.H.; Li, C.N.; Deng, N.Y. MLTSVM: A novel twin support vector machine to multi-label learning. *Pattern Recog.* **2016**, *52*, 61–74. [CrossRef]
12. Geng, X. Label distribution learning. *IEEE Trans. Knowl. Data Eng.* **2016**, *28*, 1734–1748. [CrossRef]
13. Tao, A.; Xu, N.; Geng, X. Labeling information enhancement for multi-label learning with low-rank subspace. In Proceedings of the Pacific Rim International Conference on Artificial Intelligence, Nanjing, China, 28–31 August 2018; pp. 671–683.
14. Zhang, M.L.; Zhang, Q.W.; Fang, J.P.; Li, Y.K.; Geng X. Leveraging implicit relative labeling importance information for effective multi-label learning. *IEEE Trans. Knowl. Data Eng.* **2021**, *33*, 2057–2070. [CrossRef]
15. Xu, N.; Liu, Y.P.; Geng, X. Label enhancement for label distribution learning. *IEEE Trans. Knowl. Data Eng.* **2021**, *32*, 1632–1643. [CrossRef]
16. Shao, R.F.; Xu, N.; Geng, X. Multi-label learning with label enhancement. In Proceedings of the International Conference on Data Mining, Singapore, 17–20 November 2018; pp. 437–446.
17. Yao, Y.Y. Three-way decision: An interpretation of rules in rough set theory. In Proceedings of the 4th International Conference on Rough Sets and Knowledge Technology, Gold Coast, Australia, 14–16 July 2009; pp. 642–649.
18. Yao, Y.Y. Three-way decision and granular computing. *Int. J. Approx. Reason.* **2018**, *103*, 107–123. [CrossRef]
19. Yao, Y.Y. Tri-level thinking: Models of three-way decision. *Int. J. Mach. Learn. Cybern.* **2020**, *11*, 947–959. [CrossRef]
20. Yao, Y.Y. Three-way granular computing, rough sets, and formal concept analysis. *Int. J. Approx. Reason.* **2020**, *116*, 106–125. [CrossRef]
21. Lang, G.M.; Miao, D.Q.; Fujita, H. Three-way group conflict analysis based on pythagorean fuzzy set theory. *IEEE Trans. Fuzzy Syst.* **2020**, *28*, 447–461. [CrossRef]
22. Zhan, J.M.; Jiang, H.B.; Yao, Y.Y. Three-way multiattribute decision-making based on outranking relations. *IEEE Trans. Fuzzy Syst.* **2021**, *29*, 2844–2858. [CrossRef]
23. Zhang, X.Y.; Gou, H.Y.; Lv, Z.Y.; Miao, D.Q. Double-quantitative distance measurement and classification learning based on the tri-level granular structure of neighborhood system. *Knowl.-Based Syst.* **2021**, *217*, 106799. [CrossRef]
24. Jiang, C.M.; Guo, D.D.; Sun, L.J. Effectiveness measure for TAO model of three-way decisions with interval set. *J. Intell. Syst.* **2021**, *40*, 11071–11084. [CrossRef]
25. Yang, J.L.; Yao, Y.Y.; Zhang, X.Y. A model of three-way approximation of intuitionistic fuzzy sets. *Int. J. Mach. Learn. Cybern.* **2022**, *13*, 163–174. [CrossRef]
26. Guo, D.D.; Jiang, C.M.; Wu, P. Three-way decision based on confidence level change in rough set. *Int. J. Approx. Reason.* **2022**, *143*, 57–77. [CrossRef]
27. Huang, X.F.; Zhan, J.M.; Ding, W.P.; Pedrycz, W. An error correction prediction model based on three-way decision and ensemble learning. *Int. J. Approx. Reason.* **2022**, *146*, 21–46. [CrossRef]
28. Zhang, Y.J.; Miao, D.Q.; Zhang, Z.F.; Xu, J.F.; Luo, S. A three-way selective ensemble model for multi-label classification. *Int. J. Approx. Reason.* **2018**, *103*, 394–413. [CrossRef]
29. Ren, F.J.; Wang, L. Sentiment analysis of text based on three-way decisions. *J. Intell. Fuzzy Syst.* **2017**, *33*, 245–254. [CrossRef]
30. Zhang, Y.J.; Miao, D.Q.; Pedrycz, W.; Zhao, T.N.; Xu, J.F.; Yu, Y. Granular structure-based incremental updating for multi-label classification. *Knowl.-Based Syst.* **2020**, *189*, 105066. [CrossRef]
31. Zhang, Y.J.; Zhao, T.N.; Miao, D.Q.; Pedrycz, W. Granular multilabel batch active learning with pairwise label correlation. *IEEE Trans. Syst. Man Cybern.-Syst.* **2022**, *52*, 3079–3091. [CrossRef]
32. Qian, W.B.; Huang, J.T.; Wang, Y.L.; Xie, Y.H. Label distribution feature selection for multi-label classification with rough set. *Int. J. Approx. Reason.* **2021**, *128*, 32–55. [CrossRef]
33. Kongsorot, Y.; Horata, P.; Musikawan,P.; Sunat, K. Kernel extreme learning machine based on fuzzy set theory for multi-label classification. *Int. J. Mach. Learn. Cybern.* **2019**, *10*, 979–989. [CrossRef]
34. Yuichi, O.; Naoki, M.; Yusuke, N.; Hisao, I. Multiobjective fuzzy genetics-based machine learning for multi-label classification. In Proceedings of the IEEE International Conference on Fuzzy Systems, Glasgow, UK, 19–24 July 2020. [CrossRef]
35. Che, X.Y.; Chen, D.G.; Mi, J.S. Feature distribution-based label correlation in multi-label classification. *Int. J. Mach. Learn. Cybern.* **2021**, *12*, 1705–1719. [CrossRef]

36. Xiao, F.Y. A distance measure for intuitionistic fuzzy sets and its application to pattern classification problems. *IEEE Trans. Syst. Man Cybern.-Syst.* **2021**, *51*, 3980–3992. [CrossRef]
37. Tian, Y.; Sun, M.; Deng, Z.B.; Luo, J.; Li, Y.Q. A new fuzzy set and nonkernel svm approach for mislabeled binary classification with applications. *IEEE Trans. Fuzzy Syst.* **2017**, *25*, 1536–1545. [CrossRef]
38. Tian, Y.; Deng, Z.B.; Luo, J.; Li, Y.Q. An intuitionistic fuzzy set based (SVM)-V-3 model for binary classification with mislabeled information. *Fuzzy Optim. Decis. Mak.* **2018**, *17*, 475–494. [CrossRef]
39. Rezvani, S.; Wang, X.Z.; Pourpanah, F. Intuitionistic fuzzy twin support vector machines. *IEEE Trans. Fuzzy Syst.* **2019**, *27*, 2140–2151. [CrossRef]
40. Zhang, M.L.; Wu, L. LIFT: Multi-label learning with label-specific features. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 107–120. [CrossRef]
41. Guo, Y.M.; Chung, F.L.; Li, G.Z.; Wang, J.C.; Gee, J.C. Leveraging label-specific discriminant mapping features for multi-label learning. *ACM Trans. Knowl. Discov. Data* **2019**, *13*, 24. [CrossRef]
42. Yu, Z.B.; Zhang, M.L. Multi-label classification with label-specific feature generation: A wrapped approach. *IEEE Trans. Pattern Anal. Mach. Intell.* 2021, *in press*. [CrossRef]
43. Jia, X.Y.; Lu, Y.N.; Zhang, F.W. Label enhancement by maintaining positive and negative label relation. *IEEE Trans. Knowl. Data Eng.* 2021, *in press*. [CrossRef]
44. Zheng, Q.H.; Zhu, J.H.; Tang, H.Y.; Liu, X.Y.; Li, Z.Y.; Lu, H.M. Generalized label enhancement with sample correlations. *IEEE Trans. Knowl. Data Eng.* 2021, *in press*. [CrossRef]
45. Ha, M.H.; Wang, C.; Chen, J.Q. The support vector machine based on intuitionistic fuzzy number and kernel function. *Soft Comput.* **2013**, *17*, 635–641. [CrossRef]
46. Zhang, M.L.; Li, Y.K.; Yang, H.; Liu, X.Y. Towards class-imbalance aware multi-label learning. *IEEE Trans. Cybern.* 2020, *in press*. [CrossRef] [PubMed]
47. Schapire, R.; Singer, Y. A boosting-based system for text categorization. *Mach. Learn.* **2000**, *39*, 135–168. [CrossRef]
48. Zhang, M.L.; Zhou, Z.H. ML-KNN: A lazy learning approach to multi-label learning. *Pattern Recog.* **2007**, *40*, 2038–2048. [CrossRef]
49. Zhu, Y.; Kwok, J.T.; Zhou, Z.H. Multi-label learning with global and local label correlation. *IEEE Trans. Knowl. Data Eng.* **2018**, *30*, 1081–1094. [CrossRef]
50. Wang, R.; Kwong, S.; Wang, X.; Jia, Y.H. Active *K*-Labelsets Ensemble Multi-Label Classification. *Pattern Recog.* **2021**, *109*, 107583. [CrossRef]
51. Demsar, J. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.* **2006**, *7*, 1–30.