Granular Multilabel Batch Active Learning With Pairwise Label Correlation

Yuanjian Zhang, Tianna Zhao, Duoqian Miao, and Witold Pedrycz, *Life Fellow, IEEE*

***Abstract*—Abundant data with limited labeling are a widespread bottleneck in multilabel learning. Active learning (AL) is an effective solution to gradually enhance model robustness, however how to effectively extend instance selection criteria to multilabel case remains challenging. Considering the label specificity and label correlation, a granular batch mode-based ranking active model for the multilabel (GBRAML) is proposed. Taking a bottom-up view, three granulation operators are successively constructed to formulate three granular structures. In low-level granulation operator, auxiliary label is introduced to enhance the informativeness and representativeness of each label. The contribution of labels to the usefulness of instances is incorporated with pair-wise label correlation, and is considered in the middle-level granulation operator. The labeling priority is determined by ranking the scorings coming from high-level granulation operator. To alleviate the impact of skewed label correlation, we take a three-way strategy on fitness of representative label correlation, thus a three-way GBRAML model (TGBRAML) is devised. Extensive experiments on six multilabel benchmark demonstrate GBRAML gains 5.4% and 210.1% improvement on MicroF1 and Average Precision over state-of-the-art batch mode multilabel active learning. The effectiveness of three-way decisions in multilabel AL is also verified.**

***Index Terms*—Active learning, batch mode, granular computing (GrC), label correlation, multilabel, three-way decisions.**

I. Introduction

IN MULTILABEL scenario, one example is associated with potentially dozens of labels simultaneously, and the primary goal of multilabel learning is to train a model that can determine label assignments or predict label ranking on unseen instances [1], [2]. Representative applications involve smart grid management [3], disease diagnosis [4], and image classification [5]. A vital prerequisite in learning a desirable multilabel classifier is to have plentiful information on label associations. While the speed of data collection is rapidly growing with the advances in automation, the expenditure on label annotations is increasingly intolerable. The boundary among different labels is more ambiguous than ever, thus sophisticated identification is required for the determination of presence/absence of each label.

Active learning (AL) [6] attempts to reduce the burden of manual annotation by selecting a number of valuable instances for querying based on some criteria. During the past decades, three criteria, i.e., informativeness, representativeness and diversity, have been advocated. The criterion of informativeness measures the ability of an instance to reduce the generalization error of trained model. The criterion of representativeness examines the ability of an instance to restore the distribution of unlabeled data. The criterion of diversity measures the information redundancy of an instance to classifier construction. Under the combination of the above-mentioned criteria, many multilabel AL algorithms perform in a batch-mode manner [7]–[10]. The batch is usually composed of unlabeled instances and iteratively enriches the labeled set. Typical operations in one labeling iteration include instance selection, batch annotation and model reconstruction, where AL works mainly in the instance selection stage. It is worth mentioning that almost all multilabel active algorithms assume that labels are determined by all conditions and do not discriminate the learning difficulty of each label, which yields suboptimal performance.

Granular computing (GrC) [11] is a methodology concerning the definition, transformation and computing of information granules. By adopting GrC, researchers generate a granular representation of data with the multilevel characteristic. For example, AL can be hierarchically processed on single-label data and the structural cost combining misclassification with annotation is significantly reduced [12]. However the applicability to multilabel AL has yet not been examined. Three-way decisions (3WD) [13], [14] is a triarchic theory of GrC that simulates the actions a person may take when facing uncertainty. The semantics of decision include but not limited to classification, clustering, decision support, concept learning and active learning [15]–[17]. The compatibility motivates us to explore uncertainty in learnt multilabel model.

Label correlation and class imbalance are two characteristics of multilabel problems. The label correlation measures the possibility that two or more labels co-occur or not. For

Manuscript received October 26, 2020; revised January 28, 2021; accepted February 21, 2021. Date of publication March 12, 2021; date of current version April 15, 2022. This work was supported by the National Natural Science Foundation of China under Grant 61976158, Grant 61763031, Grant 61906137, and Grant 62076182. This article was recommended by Associate Editor C.-T. Lin. *(Corresponding author: Duoqian Miao.)*

Yuanjian Zhang is with the Department of Computer Science and Technology, Tongji University, Shanghai 201804, China (e-mail: 96zhangyj@tongji.edu.cn).

Tianna Zhao and Duoqian Miao are with the Department of Computer Science and Technology, Tongji University, Shanghai 201804, China (e-mail: 1810375@tongji.edu.cn; dqmiao@tongji.edu.cn).

Witold Pedrycz is with the Department of Computer Science and Technology, Tongji University, Shanghai 201804, China, also with the Department of Electrical and Computer Engineering, Alberta University, Edmonton, AB T6R 2V4, Canada, and also with the System Research Institute, Polish Academy of Sciences, PL-01447 Warsaw, Poland (e-mail: wpedrycz@ualberta.ca).

This article has supplementary material provided by the authors and color versions of one or more figures available at https://doi.org/10.1109/TSMC.2021.3062714.

Digital Object Identifier 10.1109/TSMC.2021.3062714

instance, the label "boat" is more correlated with "sea" than "desert." The additional knowledge is unknown beforehand and can be viewed as a metric for the assessment of label complexity. Class imbalance refers to the imbalanced distribution between positive class (those who are associated with instances) and negative class (those who are not associated with instances). AL on multilabel data should be aware of these characteristics so that the overall uncertainty toward valuable instances is minimized [18]. This article formulates the batch selection as an instance ranking problem and presents a granular batch mode-based ranking active model for multilabel (GBRAML). In this model, the unlabeled instances are queried across all labels independently via low-level granulation operator, and those which have modified large separation margin, a margin that is weighted by neighborhood-based class distribution, gain the priority in the finest granule. By incorporating label complexity and fitness of label correlation, the label importance for annotation is considered via middle-level granulation operator. An ordered set for unlabeled instances is deduced via high-level granulation operator. Furthermore, by introducing three-way decisions on the judgment of fitness of label correlation, the skewed label correlation in GBRAML model is solved.

Our contributions can be summarized as follows.

1) To the best of our knowledge, this is the first effort to perform multilabel AL with the consideration of data characteristics stemming from a GrC-based perspective. We utilize the class-imbalance and label correlation at low-level and middle-level layers, respectively, to generate an effective query.
2) Two instance selection criteria, representativeness and informativeness, are simultaneously leveraged at low-level and middle-level granules. The proposed ranking strategy at high-level granule is thereby more effective.
3) We extend three-way decisions to multilabel AL and alleviate the influence of estimated label correlation bias.

The remainder of this article is organized as follows. Section II reviews some related work; Section III presents our proposed model for multilabel active learning, which is then incorporated with three-way decisions in Section IV; experimental results are reported in Section V; Section VI concludes this work by identifying future directions.

II. Related Work

Instance selection [19] is a fundamental issue in active learning. An utility function is required to evaluate the urgency of unlabeled instances. While instance uncertainty remains a challenge in multilabel setting, the versatile implementations of utility function are hierarchically defined [20]–[22]. The underlying reason is the different interests of stakeholder, and one typical style is defined as multiple scoring functions with an aggregation function [23]. The availability of data determines the AL mechanism. In this article, we focus on scenarios in which a large collection of unlabeled data and a small set of labeled data is available, i.e., pool-based active learning.

Many scholars have investigated a plethora of selection criteria for pool-based multilabel active learning. The pioneering work is reported in [24], in which two selection strategies named "max loss" (ML) and "mean ML" (MML) are put forward, respectively. The informative candidates are with the maximal expected loss decrement on most certain label. A similar work can be found in [25], where the count of associated labels is predicted by an auxiliary regression model. Instances are selected based on the principle "maximal expected loss reduction with maximal confidence" (MMC), which is implemented via the approximation between expected label count and classification probabilities. The work is further extended in [26], where the label ranking is adopted for the identification of separation margin and the expected label cardinality is examined meanwhile. By minimizing difference of data distribution in labeled and unlabeled parts, an approach which considers both representativeness and diversity is considered in [27]. The positive-prone class is more preferred, resulting in a steady rise of performance. To cope with outlier labels in images, a maximum correntropy criterion (MCC) [28] with the combination of informativeness and representativeness is proposed, in which informativeness is interpreted as minimum margin, and representativeness is defined as the consistency between labels and features. By adopting the soft Hamming loss reduction criterion [29], informative messages committed by users are preferentially responded. This criterion is composed of a Hamming loss reduction and a maximum margin reduction. Considering the good generalization of hyperplane based classifier, linear model is considered as the baseline in our model.

The expansion from serial mode (one instance per iteration) to batch mode (multiple instances per iteration) can significantly improve the utilization of manual labeling, yet information overlap among the selected batch may incur. In [30], Fisher information matrix is employed to measure the uncertainty of unlabeled instances. The selection problem is formulated via semidefinitive programming and optimized via upper bound. A batch instance selection criterion combining informativeness with diversity is presented in [31]. The notion is that instances with high informativeness measured by current classifier individually and jointly less redundancy should be annotated. However, the problem is formulated as an NP-hard integer quadratic programming. In [32], the informativeness and diversity of selected batch are balanced in the granularity of instance-label pair. The potentially high-order label correlation is identified at label level. To circumvent the drawback, some efforts on customizing problem formulation and optimization strategies have been attempted. AL completed on relative attributes for semantic understanding is considered in [33], in which a diverse expected gradient model is constructed. To guarantee the informativeness and diversity, a two-step heuristic method is proposed to iteratively generate an approximate optimal query set. The confidence of unlabeled instances between two most likely classes is utilized as a metric to measure diversity in [34], and is implemented via two techniques named "angle-based diversity" and "clustering-based diversity," respectively. Coupled with redundant instance removal strategies in subproblem level, a criterion combining informativeness and diversity is presented. This work is further improved in [35] by two adjustments in informativeness and diversity, in which the uncertain instances among all hyperplanes are refined by kernel k-means clustering. While instance redundancy is reduced dramatically, the computation on pair-wise unlabeled instances is costly.

TABLE I
CHARACTERISTICS FOR EXISTING MULTILABEL AL APPROACHES. WE USE S, B TO REPRESENT SINGLE/MULTIPLE INSTANCE SELECTION MODE, AND I, R, D TO REPRESENT THE INSTANCE SELECTION CRITERION OF INFORMATIVENESS, REPRESENTATIVENESS, AND DIVERSITY

Method	Year	Selection Type	Selection Criteria
[24]	2004	S	I
[25]	2009	S	I
[30]	2009	B	I+R
[31]	2011	B	I+D
[26]	2013	S	I+R
[32]	2014	B	I+D
[28]	2017	S	I+R
[36]	2017	B	I+R+D

Ranked batch mode active learning [36] is a flexible framework which constitutes three procedures, named as "uncertainty estimation," "ranked batch construction," and "oracle labeling." With this routine, unlabeled instances are ranked by an integrated scoring function which weighs similarity and uncertainty simultaneously. Uncertainty sampling by query-by-committee strategy is complemented by diversity and density in [37]. The optimal batch is searched in a greedy fashion based on the score ranking. In [38], multiple criteria for informativeness and diversity on unlabeled instances are used by referring to labeled instances and current model, respectively.

The characteristics of state-of-the-art multilabel AL strategies are summarized in Table I. In this article, we consider batch mode AL strategy by leveraging both informativeness and representativeness on hierarchical score ranking. The purpose of ranking is to prioritize unlabeled instances that are both informative and representative across label space. By iteratively manipulating limited annotations, the latent label associations are gradually recovered. To objectively evaluate the contribution of each label for a given instance, a hierarchical evaluation with three granulation operators is proposed. First, we examine both the pseudolabel generated by current model and local label distributions of unlabeled label. Instances with contradictory label association estimations receive higher scores. Second, fitness of representative label correlations on each label is leveraged. Instances with large bias/variance estimation with respect to selected label correlations receive higher scores. Finally, an aggregation strategy across all labels is performed. As the estimation of label correlation from limited labeled instances may be deviated [39], [40], we further employ three-way decisions on fitness evaluation, which is different from the cost-sensitive on hierarchical multilabel active learning [41]. Details regarding GBRAML and its variations will be elaborated in the following sections.

III. GBRAML MODEL

A. Notations

We present essential notations of the batch mode AL for multilabel problem. Let $D = D_l \cup D_u$ denotes a multilabel dataset which constitutes a labeled set D_l and an unlabeled set D_u. For an arbitrary instance $\mathbf{x}_i$ in D_l, it is composed of d-dimension features and m-dimension labels, denoted as

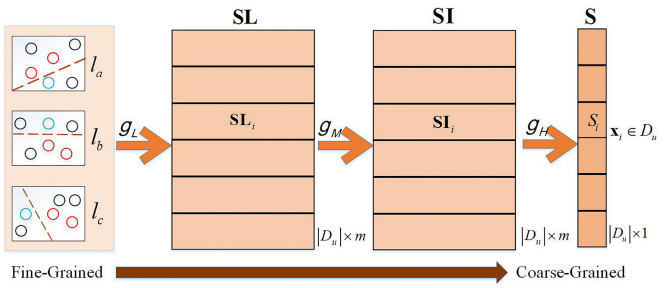

Fig. 1. Pipeline of GBRAML: we conduct AL for multilabel from left-hand side to right-hand side, transforming the fine-grained representation in each label to coarse-grained representation for labeling priority. For simplicity, we only present the distribution of six instances with respect to three labels (namely, l_a, l_b, and l_c), with red circle representing the negative class, light blue circle representing the positive class, and the remaining representing unlabeled instances. The granulation operators g_L, g_M and g_H induce the low-level, middle-level, and high-level granules, respectively. By subsequently conducting, we determine the batch of desired instances. Detailed components in g_L, g_M and g_H will be elaborated later. (Figures best viewed in color.)

$\mathbf{x}_i = [x_{i1}, x_{i2}, \ldots, x_{id}]$ and $\mathbf{y}_i = [y_{i1}, y_{i2}, \ldots, y_{im}]$, respectively. Thus, we have $D_l = \{(\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2), \ldots, (\mathbf{x}_{n_l}, \mathbf{y}_{n_l})\}$. The value of y_{ij} represents the association of the jth label to the ith instance, with $\mathbf{x}_i$ has the jth label if $y_{ij} = 1$ and otherwise if $y_{ij} = 0$. The label associations of unlabeled instances in D_u are totally unknown, i.e., $\mathbf{y}_i \in \{0, 1\}^m \quad \forall i > n_l$. The goal of the problem is to find a ranked batch $B (B \subset D_u)$ determined by $\mathbf{S}$ with the count card(B), such that the annotations on B can significantly improve the classifier $\mathbf{w} = [\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_m]$ learnt on D_l, where $\mathbf{w}_c = [w_{c1}, w_{c2}, \ldots, w_{cd}]^T$, and the included component w_{cj} corresponds to the weight of jth feature to the cth label. The vector $\mathbf{S}_{|D_u| \times 1}$ quantifies the priority of unlabeled instances. It is hierarchically constructed, with $\mathbf{SI}_{|D_u| \times m}$ and $\mathbf{SL}_{|D_u| \times m}$ representing instance-level score and label-level score, respectively. In each iteration, we generate $\mathbf{S}$ by consecutively applying three granule operators g_H, g_M and g_L, and the details will be elaborated later. The pipeline of GBRAML is shown in Fig. 1.

B. Low-Level Granule: Label-Specific Uncertainty

Given limited known labels, it is difficult to discriminate different labels via an identical feature space. To describe the characteristics of each label, the concept label-specific features [42] is proposed. With iterative reconstruction of label-dependent features, the semantics of labels are gradually clarified. Considering the reconstruction efficiency and evaluation interdependency, we seek a label-specific learning approach satisfying the following conditions.

1) Low-order correlations are leveraged in constructions of label-specific features.
2) The weaker the label correlation is, the more different in feature components becomes.

Fortunately, the method LLSF [43] is ideal for the two requirements. First, second-order correlation is considered. Second, features with higher positive/negative dependency of cth label have larger absolute value in w_{cj} with the same polarity, and $w_{cj} = 0$ holds if the jth feature is irrelevant to cth label.

TABLE II
EXAMPLE OF CONTINGENCY TABLE

	Count of C	Count of $\neg C$
Group 1	O_1	O_2
Group 2	O_3	O_4

Specifically, the formulation of LLSF is as follows:

$$\min_{\mathbf{w}_i} \frac{1}{2}\|\mathbf{X}\mathbf{w}_i - \mathbf{Y}_i\|_2^2 + \frac{\alpha}{2}\sum_{j=1}^{m} r_{ij}\mathbf{w}_i^T\mathbf{w}_j + \beta\|\mathbf{w}_i\|_1 \tag{1}$$

where $\mathbf{X}$ is the feature space of instances in D_l and $\mathbf{Y_i}$ is the vector representing the ith label. $r_{ij} = 1 - c_{ij}$, c_{ij} represents the correlation between labels y_i and y_j, which is measured by cosine similarity. $\alpha \geq 0$ and $\beta \geq 0$ are two tradeoff parameters. The problem can be optimized via accelerated proximal gradient algorithm. Then pseudolabel of an unlabeled instance $\mathbf{x}_i$ (that is $\hat{\mathbf{y}}_i$) can be assigned as

$$\hat{\mathbf{y}}_i = \text{sgn}(\mathbf{x}_i\mathbf{w} - \tau) \tag{2}$$

where $\text{sgn}(\cdot)$ is a function with the value equals to 1 for positive argument and 0 otherwise. τ is a threshold that converts the regression results to classification outcomes.

Unlike margin-based approach, which directly selects the instances with minimum separation margin, we consider the local information of $\mathbf{x}_i$ as well. k-nearest neighbor is a straightforward solution in describing local information, however neither a large k value nor a small k value is appropriate for the estimation of label associations. A neighbor induced by a small k is confused by the imbalanced label associations, whereas a neighbor induced by a large k is degenerated as lack of specificity. Instead of finding an optimal k, we address it by defining the super k-nearest neighborhood of $\mathbf{x}_i$ on cth label [that is, $SN_k^c(\mathbf{x}_i)$, see Fig. 2] as

$$SN_k^c(\mathbf{x}_i) = \bigcup\left\{N_k^c(\mathbf{x}_h)\middle|\mathbf{x}_i \in N_k^c(\mathbf{x}_h)\right\} \tag{3}$$

where $N_k^c(\mathbf{x}_h)$ is the neighbors of instance $\mathbf{x}_h$ constructed by the label-specific features on cth label. Regarding $SN_k^c(\mathbf{x}_i)$ as a local representation of $\mathbf{x}_i$, we estimate the auxiliary label of $\mathbf{x}_i$ (that is $\tilde{\mathbf{y}}_i$) via chi-square test on contingency table [44]. The result of chi-square test reveals whether there is statistical difference between two groups of observations, and the chi-square value is calculated as

$$\chi^2 = \sum_{i,j}\frac{\left(O_{i,j} - E_{i,j}\right)^2}{E_{i,j}} \tag{4}$$

where $O_{i,j}$ and $E_{i,j}$ are the observed frequencies and expected frequencies that generated by a contingency table. Table II shows a 2×2 contingency table, where C and $\neg C$ denote the events that instances are associated with the cth label or not. Group 1 and Group 2 represent the label association in D_l on cth label and pseudo label association within neighborhoods generated by $\mathbf{x}_h$(i.e., $N_k^c(\mathbf{x}_h)$) on the cth label, respectively. The value χ^2, denoted as χ_h^2, is computed as

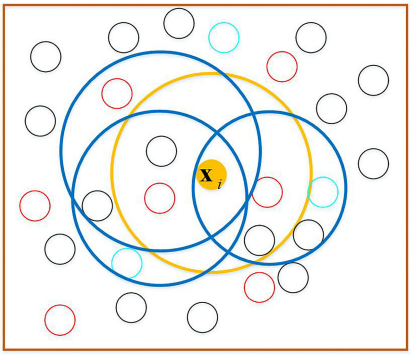

Fig. 2. Toy example of the generation of $SN_k^c(\mathbf{x}_i)$ (k=5): Given an unlabeled instance $\mathbf{x}_i$ (orange spot), the $SN_k^c(\mathbf{x}_i)$ includes not only the neighborhood $N_k^c(\mathbf{x}_i)$ identified by a distance metric on label-specific features (the orange circle), but also the neighborhoods containing $\mathbf{x}_i$ (the dark blue circle). The red spots, light blue spots, and black spots denote negative class, positive class, and unlabeled with respect to cth label, respectively. (Figures best viewed in color.)

TABLE III
EXAMPLE OF ENTROPY-BASED CONTINGENCY TABLE

	Count of C	Count of $\neg C$
Group 1	$-\frac{O_1}{O_1+O_2} \times \log\frac{O_1}{O_1+O_2}$	$-\frac{O_2}{O_1+O_2} \times \log\frac{O_2}{O_1+O_2}$
Group 2	$-\frac{O_3}{O_3+O_4} \times \log\frac{O_3}{O_3+O_4}$	$-\frac{O_4}{O_3+O_4} \times \log\frac{O_4}{O_3+O_4}$

$$\chi_h^2 = \frac{\left(O_1 - CG_{1,1}\right)^2}{CG_{1,1}} + \frac{\left(O_2 - CG_{1,2}\right)^2}{CG_{1,2}} + \frac{\left(O_3 - CG_{2,1}\right)^2}{CG_{2,1}} + \frac{\left(O_4 - CG_{2,2}\right)^2}{CG_{2,2}} \tag{5}$$

where $CG_{1,1} = ([(O_1 + O_3) \times (O_1 + O_2)]/[O_1 + O_2 + O_3 + O_4])$, $CG_{1,2} = ([(O_2 + O_4) \times (O_1 + O_2)]/[O_1 + O_2 + O_3 + O_4])$, $CG_{2,1} = ([(O_1 + O_3) \times (O_3 + O_4)]/[O_1 + O_2 + O_3 + O_4])$, and $CG_{2,2} = ([(O_3 + O_4) \times (O_2 + O_4)]/[O_1 + O_2 + O_3 + O_4])$. The null hypothesis (denoted as H_0) for Table II states that two groups of class distributions are statistically indistinguishable, and it is rejected with 95% confidence if the chi-square value is larger than 3.84.[1]

Our observation is that the unlabeled instance $\mathbf{x}_i$ is likely to have the cth label if it is more positive prone as compared with prior label distribution in D_l, and *vice versa*. Recall that the scale difference between labeled instances and instance neighborhood may deteriorate the sensitivity of chi-test, we replace with the information entropy of each item in Table II and take a query-by-committee strategy across all included k-nearest neighborhoods. The generic elements in Table II are reorganized in Table III as follows.

Let $\tilde{y}_{ic}$ denotes the auxiliary label of $\mathbf{x}_i$ on cth label. It can be determined via the following expression:

$$\tilde{y}_{ic} = \text{sgn}\left(\frac{\sum_h\left(\chi_h^2 > 3.84 \wedge P\left(\hat{y}_{*c} = 1, \mathbf{x}_{*c} \in N_k^c(\mathbf{x}_h)\right) \geq \theta_c\right)}{\text{card}\left(\bigcup h\right)} \geq \theta_c\right) \tag{6}$$

[1]In a 2×2 contingency table, the degree of freedom $df = (2-1)\times(2-1) = 1$, and $P(\chi^2 \geq 0.05) = 3.84$ holds for χ^2 distribution.

Algorithm 1: ScoreInstanceByLabel

Input: Labeled set D_l, Unlabelled set D_u
Output: Label-based score matrix **SL**, linear classifier **w**

Train a classifier **w** on D_l.
for $h = 1$ *to* $card(D)$ **do**
  Compute $\hat{\mathbf{y}}_h$ as described in (2);
  for $c = 1$ *to* m **do**
    Generate $N_k^c(\mathbf{x}_h)$;
    Compute $P(\hat{y}_{*c} = 1, \mathbf{x}_{*c} \in N_k^c(\mathbf{x}_h))$;
    for $\boldsymbol{x}_i \in D_u$ **do**
      if $\boldsymbol{x}_i \in N_k^c(\boldsymbol{x}_h)$ **then**
        Substituting items in Table III into (5).
      end
    end
    for $i = 1$ *to* $card(D_u)$ **do**
      for $c = 1$ *to* m **do**
        Compute $\tilde{y}_{ic}$ as described in (6);
      end
      Compute $\mathbf{SL}_i$ as described in (7) and (8);
    end
  end
end

where $P(\hat{y}_{*c} = 1,\ \mathbf{x}_{*c} \in N_k^c(\mathbf{x}_h)) = ([\text{card}\ (\{\hat{y}_{*c} = 1 | \mathbf{x}_{*c} \in N_k^c(\mathbf{x}_h)\}]/k)$, $\theta_c = ([\text{card}(\{y_{rc} = 1 | \mathbf{x}_r \in D_l\})]/[\text{card}(D_l)])$ represents the prior probability of instances having cth label on D_l. $\text{card}(\bigcup h)$ denotes the cardinality of all h satisfying $N_k^c(\mathbf{x}_h) \subseteq SN_k^c(\mathbf{x}_i)$. Note here we do not necessarily compute $SN_k^c(\mathbf{x}_i)$, as the evaluation is an ensemble of the included neighbors.

Having information from both global perspective (i.e., $\hat{y}_{ic}$) and local perspective (i.e., $\tilde{y}_{ic}$), we consider the construction of g_L. The pseudolabel $\hat{y}_{ic}$ provides the best generalization based on known label assignments, whereas the auxiliary label $\tilde{y}_{ic}$ offers the maximal possibility of label preference based on a neighborhood structure. This implies that the annotation of $\hat{y}_{ic}$ is plausible if the labelings between $\tilde{y}_{ic}$ and $\hat{y}_{ic}$ are unanimous. Thus, for cth label, $\mathbf{x}_i$ is thus less informative (see Fig. 3). Furthermore, smaller separation margin induced by $|\mathbf{w}_c \mathbf{x}_i - \tau|$ implies larger possibility of misclassification. Without loss of generality, a granulation operator at low-level, that is g_L, is pertinent to three components, including pseudolabel ($\hat{\mathbf{y}}_i$), auxiliary label ($\tilde{\mathbf{y}}_i$), and model parameters ($\mathbf{w}$), denoted as $g_L(\hat{\mathbf{y}}_i, \tilde{\mathbf{y}}_i, \mathbf{w})$. It is defined as

$$g_L\big(\hat{\mathbf{y}}_i, \tilde{\mathbf{y}}_i, \mathbf{w}\big) = \big(\hat{\mathbf{y}}_i \oplus \tilde{\mathbf{y}}_i\big) \circ (\mathbf{1} - \text{LI}(\mathbf{x}_i \mathbf{w} - \tau)\text{LI}) \tag{7}$$

where $\text{LI} \cdot \text{LI}$ is a normalization operation which maps the input to the interval $[0, 1]$. The deduced granule structure for each unlabeled instance $\mathbf{x}_i$, that is $\mathbf{SL}_i$, is denoted as

$$\mathbf{SL}_i \triangleq g_L\big(\hat{\mathbf{y}}_i, \tilde{\mathbf{y}}_i, \mathbf{w}\big). \tag{8}$$

The normalized margin guarantees a fair comparison across different labels. The larger the value of SL_{ic} is, the more informative the $\mathbf{x}_i$ becomes. We summarize the low-level granule with Algorithm 1 for the solving of label level score **SL**. The complexity of Algorithm 1 is $O(|D|^2 d' m)$, where d' ($d' < d$) represents the average count of label-specific features among all labels.

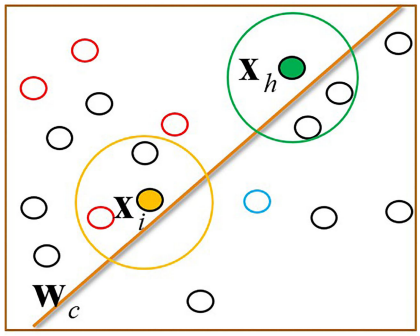

Fig. 3. Toy example of score ranking at low-level with $k = 3$. The black spots are unlabeled instances, whereas the red and blue spots represent the instances that have/donot have the label, respectively. The bold line is the hyperplane from linear classifier $\mathbf{w}_c$. The yellow circle and green circle correspond the k-nearest neighborhood of $\mathbf{x}_i$ and $\mathbf{x}_h$, respectively. Note that although $\mathbf{x}_i$ has smaller margin than $\mathbf{x}_h$, $SL_{hc} > SL_{ic}$ holds, as $\hat{y}_{ic}$ and $\tilde{y}_{ic}$ is identical, whereas $\hat{y}_{hc}$ and $\tilde{y}_{hc}$ is different. Consequently, for cth label, $\mathbf{x}_h$ is more informative than $\mathbf{x}_i$. (Figures best viewed in color.)

C. Middle-Level Granule: Label Weighting

A critical issue in (7) is the absence of label importance, which yields to the equally contribution of each label to the priority of instance annotation. This can be solved by introducing label correlation. With stronger label correlation, conducting accurate classification is easier, and thus the necessity of labeling on such instances are limited. If, however, the label correlation is negligible, it becomes necessary to annotate manually. For computational simplicity, we employ second-order label correlations on each label with most positive and most negative correlated labels. Inspired by [45], the label correlation is measured by Pearson correlation [46]. Let $\hat{\mathbf{Y}}_c$ and $\hat{\mathbf{Y}}_r$ denote the pseudolabel vectors of $\mathbf{X}$ on cth label and rth label, respectively. Then the Pearson correlation coefficient $\text{Corr}(\hat{\mathbf{Y}}_c, \hat{\mathbf{Y}}_r)$ is defined as

$$\text{Corr}\Big(\hat{\mathbf{Y}}_c, \hat{\mathbf{Y}}_r\Big) = \frac{E\Big[\Big(\hat{\mathbf{Y}}_c - \mu_{\hat{\mathbf{Y}}_c}\Big)\Big(\hat{\mathbf{Y}}_r - \mu_{\hat{\mathbf{Y}}_r}\Big)\Big]}{\lambda} \tag{9}$$

where $\lambda = \sqrt{\sum_{i=1}^{\text{card}(D)} (\hat{\mathbf{Y}}_{ic} - \mu_{\hat{\mathbf{Y}}_c})^2} \sqrt{\sum_{i=1}^{\text{card}(D)} (\hat{\mathbf{Y}}_{ir} - \mu_{\hat{\mathbf{Y}}_r})^2}$, $E[\cdot]$ denotes the variable expectation, $\mu_{\hat{\mathbf{Y}}_c}$ and $\mu_{\hat{\mathbf{Y}}_r}$ are the mean value of $\mathbf{Y}_c$ and $\mathbf{Y}_r$, respectively. $\text{Corr}(\hat{\mathbf{Y}}_c, \hat{\mathbf{Y}}_r) \in [-1, 1]$, with -1 if cth label and rth label are strongest negative linear correlation and 1 if cth label and rth label are strongest positive linear correlation. As labeling difficulty of a particular label is inversely correlated to label correlation, we formulate the difficulty of positive correlated annotation on cth label as

$$\text{Diff}_c^+ = 1 - \text{Corr}\Big(\hat{\mathbf{Y}}_c, \hat{\mathbf{Y}}_{c^+}\Big) \tag{10}$$

where $c^+ = \left\{ \arg\max_r \text{Corr}(\hat{\mathbf{Y}}_c, \hat{\mathbf{Y}}_r) \middle| \text{Corr}(\hat{\mathbf{Y}}_c, \hat{\mathbf{Y}}_r) > 0 \right\}$.

Analogously, the difficulty of negative correlated annotation on cth label can be estimated as

$$\text{Diff}_c^- = 1 + \text{Corr}\Big(\hat{\mathbf{Y}}_c, \hat{\mathbf{Y}}_{c^-}\Big) \tag{11}$$

where $c^- = \left\{ \arg\min_r \text{Corr}(\hat{\mathbf{Y}}_c, \hat{\mathbf{Y}}_r) \middle| \text{Corr}(\hat{\mathbf{Y}}_c, \hat{\mathbf{Y}}_r) < 0 \right\}$.

It is assumed in [43] that the similarity between coefficient vectors will be larger if two labels are strongly correlated. This

motivates us to examine the similarity between two classifiers (i.e., $\mathbf{w}_c$ versus $\mathbf{w}_r$) and the corresponding label correlations (i.e., $\hat{\mathbf{Y}}_c$ versus $\hat{\mathbf{Y}}_r$). We term the similarity as the fitness of model parameters with respect to label correlation. Let f_c^+ and f_c^- denote the fitness of correlation on cth label with most positive/negative correlated label, to the correlation on corresponding model parameters, respectively. They are defined as follows:

$$f_c^+ = \left|\text{Corr}(\mathbf{w}_c, \mathbf{w}_{c^+}) - \text{Corr}\left(\hat{\mathbf{Y}}_c, \hat{\mathbf{Y}}_{c^+}\right)\right| \tag{12}$$

where $c^+ = \left\{\arg\max_r \text{Corr}\left(\hat{\mathbf{Y}}_c, \hat{\mathbf{Y}}_r\right) \middle| \text{Corr}\left(\hat{\mathbf{Y}}_c, \hat{\mathbf{Y}}_r\right) > 0\right\}$

$$f_c^- = \left|\text{Corr}(\mathbf{w}_c, \mathbf{w}_{c^-}) - \text{Corr}\left(\hat{\mathbf{Y}}_c, \hat{\mathbf{Y}}_{c^-}\right)\right| \tag{13}$$

where $c^- = \left\{\arg\min_r \text{Corr}\left(\hat{\mathbf{Y}}_c, \hat{\mathbf{Y}}_r\right) \middle| \text{Corr}\left(\hat{\mathbf{Y}}_c, \hat{\mathbf{Y}}_r\right) < 0\right\}$.

A larger value on either f_c^+ or f_c^- implies that the label correlation is improperly learnt, and labeling on such instances becomes valuable. Treating the most positive label correlation as a reference, the instances receive more attention from experts if they are associated with one label but not associated with another. For the most negative correlated label, the instances receive with more attentions from experts if they are either associated or not associated with two labels simultaneously. Therefore, the granulation operator at middle level (that is g_M), is pertinent to three components, including annotation difficulty **Diff**, correlation fitness **f** and label-level score $\mathbf{SL}_i$, denoted as $g_M(\mathbf{Diff}, \mathbf{f}, \mathbf{SL}_i)$. It is defined as

$$\begin{aligned} g_M(\mathbf{Diff}, \mathbf{f}, \mathbf{SL}_i) = &\left(\mathbf{1} + \hat{\mathbf{y}}_i \oplus \hat{\mathbf{y}}_i^+\right) \circ \mathbf{SL}_i \circ e^{\mathbf{Diff}^+ \circ \mathbf{f}^+} \\ &+ \left(\mathbf{1} + \hat{\mathbf{y}}_i \odot \hat{\mathbf{y}}_i^-\right) \circ \mathbf{SL}_i \circ e^{\mathbf{Diff}^- \circ \mathbf{f}^-} \end{aligned} \tag{14}$$

where $\hat{\mathbf{y}}_i^+ = (\hat{y}_{ir})_{1\times m}$, $\hat{y}_{ir} = \arg\max_r \text{Corr}(\hat{\mathbf{Y}}_c, \hat{\mathbf{Y}}_r)$, $\hat{\mathbf{y}}_i^- = (\hat{y}_{is})_{1\times m}$, $\hat{y}_{is} = \arg\min_s \text{Corr}(\hat{\mathbf{Y}}_c, \hat{\mathbf{Y}}_s)$, $\mathbf{Diff}^+ = (\text{Diff}_c^+)_{1\times m}$, $\mathbf{Diff}^- = (\text{Diff}_c^-)_{1\times m}$, $\mathbf{f}^+ = (f_c^+)_{1\times m}$, $\mathbf{f}^- = (f_c^-)_{1\times m}$, $1 \le c \le m$. The deduced granule structure for each unlabeled instance $\mathbf{x}_i$, that is $\mathbf{SI}_i$, is denoted as

$$\mathbf{SI}_i \triangleq g_M(\mathbf{Diff}, \mathbf{f}, \mathbf{SL}_i). \tag{15}$$

We summarize the introduction of middle-level granule with Algorithm 2 for solving of instance level score **SI** and final score **S**. The complexity of Algorithm 2 is $O(\text{card}(D_u)d'm^2)$, where d' ($d' < d$) represents the average count of label-specific features among all labels.

D. High-Level Granule: Labeling Priority

To evaluate the priority of instance $\mathbf{x}_i$, we obtain the score of an unlabeled instance $\mathbf{x}_i$ by defining a granulation operator g_H. It is pertinent to instance level score $\mathbf{SI}_i$ and denoted as

$$g_H(\mathbf{SI}_i) = \sum_{c=1}^{m} SI_{ic}. \tag{16}$$

The deduced granule structure for each unlabeled instance $\mathbf{x}_i$, that is S_i, is denoted as

$$S_i \triangleq g_H(\mathbf{SI}_i). \tag{17}$$

Algorithm 2: ScoreByInstance

Input: Unlabelled set D_u, linear classifier **w**, Label-based score matrix **SL**
Output: Instance-based score matrix **SI**.

```
for i = 1 to card(D_u) do
    for c = 1 to m do
        Compute Corr(Ŷ_c, Ŷ_{c+}) as described in (9);
        Compute Diff_c^+ as described in (10);
        Compute f_c^+ as described in (12);
        Compute Diff_c^- as described in (11);
        Compute f_c^- as described in (13);
    end
    Compute SI_i as described in (14) and (15).
end
```

Algorithm 3: GBRAML

Input: Labeled set D_l, Unlabelled set D_u, Batch size $card(B)$
Output: Labeled set with new labels D_l, Unlabelled set without selected labels D_u

```
repeat
    (SL,w)=ScoreInstanceByLabel(D_u);
    SI=ScoreByInstance(D_l,D_u,w,SL);
    for x_i in D_u do
        Compute S_i as described in (16) and (17);
        Sort x_i in descending order based on S_i;
    end
    Select the top card(B) instances as B;
    Label the instances in B with ground-truth instance-label pair;
    D_l = D_l ∪ B;
    D_u = D_u − B;
until maximal iteration count reaches;
```

The larger the value of S_i is, the higher the labeling priority the $\mathbf{x}_i$ becomes.

E. Complexity Analysis

GBRAML is a combination of "ScoreInstanceByLabel" (Algorithm 1) and "ScoreByInstance" (Algorithm 2) with some necessary operations on instance ranking, batch annotation and model reconstruction.[2] The pseudocode of GBRAML is described in Algorithm 3. As the instance count is far greater than label count, the complexity of GBRAML is $O(|D|^2 d'm)$, where d' ($d' < d$) represents the average count of label-specific features among all labels.

IV. TGBRAML: Three-Way-Based GBRAML

Although the fitness of model parameter with respect to label correlation is conducive to approximating latent label correlation [see (12) and (13)], some unnecessary features (measured by **w**) may deteriorate the robustness of label classification. For example, the label "desk" may be correlated with "chair" given limited labeled paintings, however the importance of feature "round" may be unexpectedly increased, which does not offer meaningful label correlation for later classification. It is also not appropriate to overemphasize the

[2]We assume that expert is capable of labeling all labels, and the ground-truth of each selected instances is directly applied.

influence of most positive/negative correlations so that instance priority changes heavily with a small fluctuation [see (14)].

Regarding the label correlation from most positive/negative as a baseline, it is natural to divide the fitness of corresponding label-specific coefficient correlation into three cases, i.e., over-correlated, proper-correlated, and under-correlated. The tripartition is a realization of the three-way decision from a relative value view [47]. Our purpose is to design a three-way-based weighting schema on the fitness of label correlation, so that for the same instance, labels with under-correlated are more likely to be annotated than labels with proper-correlated, whereas labels with overcorrelated are least likely to be annotated. Concretely, the positive/negative consistency between label correlation of cth label to most positive/negative correlated label correlation and the correlation of $\mathbf{w}_c$ to corresponding classifiers (denoted as tf_c^+ and tf_c^-) are redefined as

$$tf_c^+ = \begin{cases} e, & \text{Corr}\left(\hat{\mathbf{Y}}_c, \hat{\mathbf{Y}}_{c^+}\right) - \text{Corr}(\mathbf{w}_c, \mathbf{w}_{c^+}) > \epsilon \\ 1, & \left|\text{Corr}\left(\hat{\mathbf{Y}}_c, \hat{\mathbf{Y}}_{c^+}\right) - \text{Corr}(\mathbf{w}_c, \mathbf{w}_{c^+})\right| < \epsilon \\ \frac{1}{e}, & \text{otherwise} \end{cases} \tag{18}$$

where $c^+ = \left\{\arg\max_r \text{Corr}(\hat{\mathbf{Y}}_c, \hat{\mathbf{Y}}_r) \middle| \text{Corr}(\hat{\mathbf{Y}}_c, \hat{\mathbf{Y}}_r) > 0\right\}$ and $\epsilon \to 0_+$. For tf_c^+, the annotation priority with under-correlated, proper-correlated, and overcorrelated are quantified as e, 1, and $(1/e)$, respectively

$$tf_c^- = \begin{cases} \frac{1}{e}, & \text{Corr}\left(\hat{\mathbf{Y}}_c, \hat{\mathbf{Y}}_{c^-}\right) - \text{Corr}(\mathbf{w}_c, \mathbf{w}_{c^-}) > \epsilon \\ 1, & \left|\text{Corr}\left(\hat{\mathbf{Y}}_c, \hat{\mathbf{Y}}_{c^-}\right) - \text{Corr}(\mathbf{w}_c, \mathbf{w}_{c^-})\right| < \epsilon \\ e, & \text{otherwise} \end{cases} \tag{19}$$

where $c^- = \{\arg\min_r \text{Corr}(\hat{\mathbf{Y}}_c, \hat{\mathbf{Y}}_r) | \text{Corr}(\hat{\mathbf{Y}}_c, \hat{\mathbf{Y}}_r) < 0\}$ and $\epsilon \to 0_+$. For tf_c^-, the annotation priority with overcorrelated, proper-correlated, and under-correlated are quantified as $(1/e)$, 1, and e, respectively.

Compared with (12) and (13), the renewed (18) and (19) take a qualitative and discriminative weighting strategy. One can infer that the contributions of two different unlabeled instances $\mathbf{x}_i$ and $\mathbf{x}_h$ from the cth component (that is, s_{ic} and s_{hc}) are similar if they have 1) similar scores at label level, and 2) similar fitness toward most positive/negative label correlation. By replacing $\mathbf{f}$ with $\mathbf{tf}$, the granulation operator $g_M(\mathbf{Diff}, \mathbf{tf}, \mathbf{SL}_i)$ (see (14)) is renewed as

$$\begin{aligned} g_M(\mathbf{Diff}, \mathbf{tf}, \mathbf{SL}_i) = &\left(\mathbf{1} + \hat{\mathbf{y}}_i \oplus \hat{\mathbf{y}}_i^+\right) \circ \mathbf{SL}_i \circ e^{\mathbf{Diff}^+ \circ \mathbf{tf}^+} \\ &+ \left(\mathbf{1} + \hat{\mathbf{y}}_i \odot \hat{\mathbf{y}}_i^-\right) \circ \mathbf{SL}_i \circ e^{\mathbf{Diff}^- \circ \mathbf{tf}^-} \end{aligned} \tag{20}$$

where $\hat{\mathbf{y}}_i^+ = (\hat{y}_{ir})_{1\times m}$, $\hat{y}_{ir} = \arg\max_r \text{Corr}(\hat{\mathbf{Y}}_c, \hat{\mathbf{Y}}_r)$, $\hat{\mathbf{y}}_i^- = (\hat{y}_{is})_{1\times m}$, $\hat{y}_{is} = \arg\min_s \text{Corr}(\hat{\mathbf{Y}}_c, \hat{\mathbf{Y}}_s)$, $\mathbf{Diff}^+ = (\text{Diff}_c^+)_{1\times m}$, $\mathbf{Diff}^- = (\text{Diff}_c^-)_{1\times m}$, $\mathbf{tf}^+ = (tf_c^+)_{1\times m}$, $\mathbf{tf}^- = (tf_c^-)_{1\times m}$, $1 \le c \le m$.

The deduced granule structure for each unlabeled instance $\mathbf{x}_i$, that is $\mathbf{SI}_i$, is denoted as

$$\mathbf{SI}_i \triangleq g_M(\mathbf{Diff}, \mathbf{tf}, \mathbf{SL}_i). \tag{21}$$

The three-way GBRAML algorithm is very much similar as GBRAML except for the evaluation of $\mathbf{SI}$. The complexity of TGBRAML is $O(\text{card}(D)^2 d' m)$, where $d' < d$ denotes the average count of label-specific feature dimension.

V. Experiments

A. Experimental Settings

We conduct two groups of experiments. For fair comparisons, the batch size card(B) is fixed as 10. The count for the labeled set is fixed as 5% of instances, and the random partition is repeated 10 times. The k-nearest neighborhood for an arbitrary instance $\mathbf{x}_i$, $N_k^c(\mathbf{x}_i)$, is generated via Euclidean distance. All AL algorithms continue until the size of labeled instances amounts to 80% of instances. The goal for the first group is to validate the effectiveness of GBRAML. To comprehensively examine the performance of GBRAML, the experiments are not merely compared with some state-of-the-art approaches, but also with some variations of GBRAML. Due to the computational efficiency and label-specific learning mechanism, we adopt LLSF [43] as a baseline of GBRAML, with recommended settings $\alpha = 2^8$, $\beta = 2^4$, and $\tau = 0.5$. Detailed settings of algorithms are given as.

1) *BMAL [31]:* Batch mode AL selects a batch of instances that simultaneously maximizes the high uncertainty and pairwise instance divergence. Support vector machine (SVM) with a polynomial kernel of degree 2 is applied for model reconstruction.
2) *Adaptive [26]*[3]*:* We select the most informative batch by ranking overall results from both max-margin prediction uncertainty and label cardinality inconsistency. The two perspectives are adaptively integrated, with the adaptive parameter β searched in $\{0, 0.1, \ldots, 1\}$. The optimal β is determined if the approximate generalization error reaches minimum. SVM with a Gaussian kernel of degree 2 is applied and tradeoff parameter $C = 100$.
3) *BatchRank [10]:* This method is an efficient version of *BMAL* and takes a ranking formulation view, with the tradeoff parameter $\lambda = 100$. Logistic regression (LR) is applied for model reconstruction.
4) *Random:* This method randomly selects instances and reconstructs model via LLSF.
5) *MGBRAML:* Mean fusion of granular batch ranking for multilabel. The scores of instances are generated via the sum of the low-level granule on each label (i.e., $S(\mathbf{x}_i) = \sum_c \text{SL}_{ic}$), without the consideration of middle-level granule. The classification, as well as the optimization of $\mathbf{w}$, is learnt via LLSF.
6) *GBRAML_LR:* Granular batch mode-based ranking active model for multilabel with logistic regression. The classification, as well as the optimization of $\mathbf{w}$, is learnt via LLSF. The neighborhood size k is fixed as 10.
7) *GBRAML_SVM:* Granular batch mode-based ranking active model for multilabel with support vector machine. The nonzero label-specific coefficient matrix $\mathbf{w}$ learnt from LLSF is utilized to train multiple binary classifiers with binary relevance support vector machine.

[3]Source code: https://carleton.ca/scs/people/yuhong-guo/

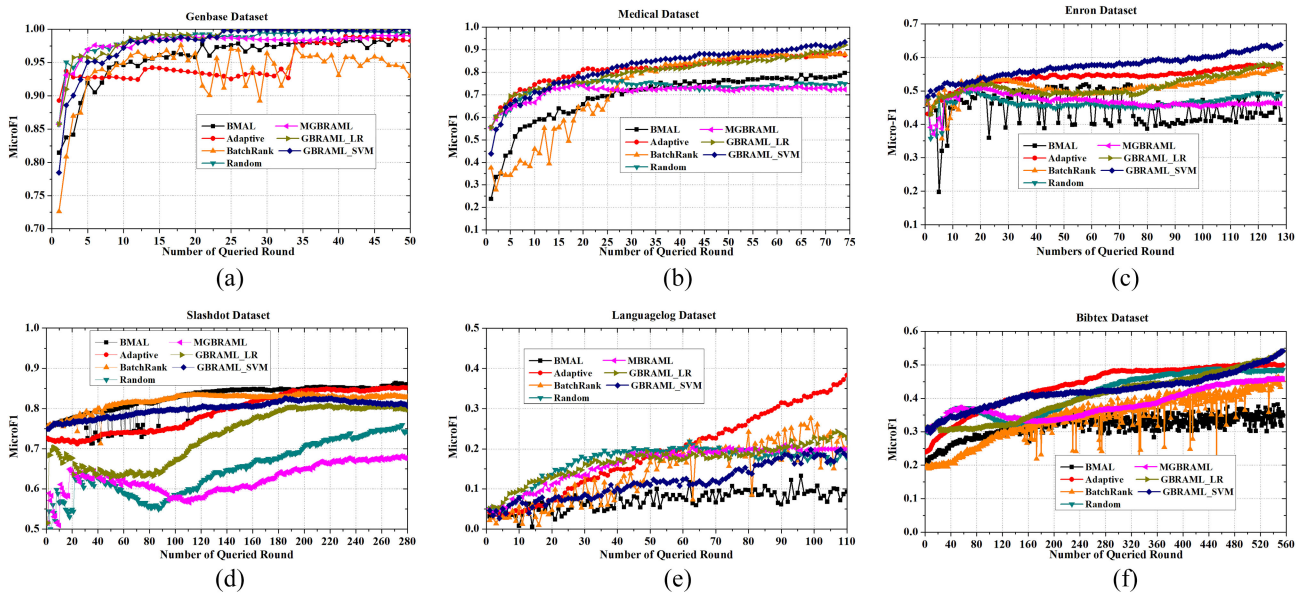

Fig. 4. Comparison of AL in multilabel datasets on MicroF1. (Figures best viewed in color.) (a) Genbase. (b) Medical. (c) Enron. (d) Slashdot. (e) Languagelog. (f) Bibtex.

TABLE IV
CHARACTERISTICS OF DATA SETS

Data	# Instances	# Features	# Labels	# Cardinality
Genbase	662	1185	27	1.252
Medical	978	1449	45	1.245
Enron	1702	1001	53	3.39
Slashdot	3782	1079	22	1.18
LanguageLog	1460	1004	75	1.18
Bibtex	7395	1836	159	2.402

For simplicity, the linear kernel is implemented. The neighborhood size k is fixed as 10.

The second experiment evaluates the performance of TGBRAML, with two variations of TGBRAML named as TGBRAML_LR and TGBRAML_SVM, respectively. For all TGBRAML variations, $\epsilon = 10^{-5}$. By comparing with GBRAML, we examine the effectiveness of three-way decisions on multilabel active learning.

All experiments are performed on six benchmarks, the details of which are summarized in Table IV. In Table IV, 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. The comparisons are examined in frequently considered benchmark in Mulan[4] and Meka.[5] All experiments are coded in MATLAB 2017b and completed on a workstation with the following specification: Intel Core12 i7-6800K 3.40GHz CPU, 64GB of memory with 64-bit ubuntu 16.0.4 operation system.The classification performance is evaluated using Micro F_1-measure (abbreviated as MicroF1) and Average Precision (abbreviated as AP) [1]. For both metrics, the larger values are, the better the performance becomes. We repeat all considered methods for five times, and compare the average performance.

[4]http://mulan.sourceforge.net/datasets.html
[5]http://meka.sourceforge.net

B. Results

1) Comparison With State-of-the-Art Methods: Figs. 4 and 5 report the average performance on MicroF1 and AP across the benchmarks, respectively. In each subfigure, the x-axis denotes the round of active query and the y-axis denotes the performance (i.e., MicroF1 and AP) obtained on the unlabeled set.

For MicroF1, we observe from Fig. 4(a)–(f) that although the absolute performance is increasing in general as more instances are labeled, the relative superiority of GBRAML (that is, GBRAML_LR and GBRAML_SVM) to other state-of-the-art algorithms are different. In all cases except dataset "Languagelog," GBRAML_SVM performs better than GBRAML_LR on average. The difference is particularly significant when the count of selected instances is limited (see round before 120 for "Slashdot" and round before 100 for "Bibtex"). It suggests that the application of the SVM can compensate for the shortcomings of biased label distribution. With a linear kernel, the performance of GBRAML_SVM can be at least comparable to the state-of-the-art algorithms, and is superior on dataset "Genbase," "Medical," and "Enron." This observation demonstrates the informativeness of selected nonzero features and instances. A more in-depth observation shows that the improvement from *Random* or MGBRAML to GBRAML_LR is larger than the advances from GBRAML_LR to GBRAML_SVM, which implies the contribution of instance selection is more significant. The functionality of middle-level granule is more important than low-level granule, as the performance on MicroF1 between *Random* and MGBRAML is similar and in most cases, worse than GBRAML_LR among all datasets.

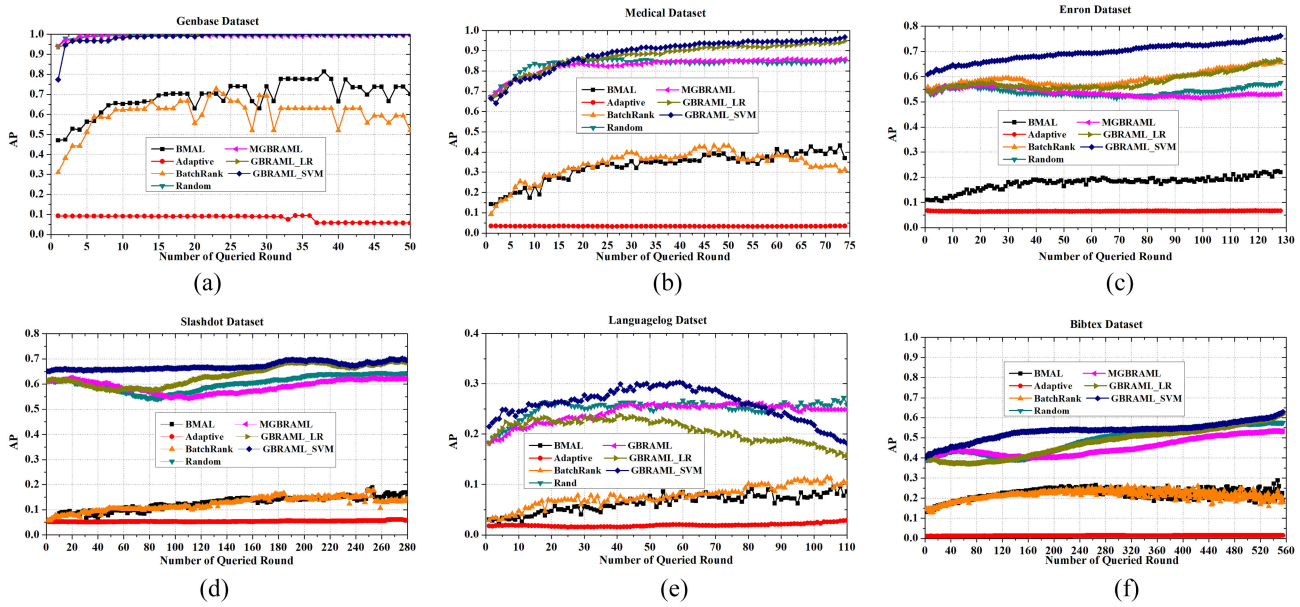

Fig. 5. Comparison of AL in multilabel datasets on AP. (Figures best viewed in color.) (a) Genbase. (b) Medical. (c) Enron. (d) Slashdot. (e) Languagelog. (f) Bibtex.

TABLE V
EXPERIMENTAL RESULTS OF EACH COMPARISON ALGORITHM (MEAN±STD WITH ALGORITHM RANKING FOLLOWED AND SUMMARIZED) ON MICROF1

Data set	Algorithms	Number of Queried Round (percentage of the unlabeled data)						
		5%	10%	20%	30%	40%	50%	80%
Genbase	BMAL	0.8657±0.366(6)	0.9150±0.198(7)	0.9501±0.046(6)	0.9606±0.028(6)	0.9721±0.007(5)	0.9741±0.004(5)	0.9872±0.008(5)
	Adaptive	0.9225±0.171(4)	0.9276±0.089(5)	0.9316±0.086(7)	0.9374±0.019(7)	0.9294±0.003(7)	0.9338±0.004(6)	0.9844±0.002(6)
	BatchRank	0.8449±0.079(7)	0.9221±0.027(6)	0.9580±0.006(5)	0.9634±0.009(5)	0.9417±0.033(6)	0.9322±0.028(7)	0.9441±0.009(7)
	Random	**0.9349±0.144**(1)	0.9670±0.070(2)	0.9839±0.037(2)	0.9900±0.002(2)	0.9897±0.002(3)	0.9930±0.002(3)	0.9954±0.001(3)
	MGBRAML	0.9295±0.143(2)	**0.9698±0.086**(1)	0.9787±0.052(4)	0.9863±0.002(3)	0.9868±0.001(4)	0.9845±0.001(4)	0.9902±0.000(4)
	GBRAML_LR	0.9285±0.045(3)	0.9591±0.035(3)	**0.9860±0.005**(1)	**0.9904±0.003**(1)	**0.9958±0.005**(1)	**1.0000±0.000**(1)	**1.0000±0.000**(1)
	GBRAML_SVM	0.8892±0.063(5)	0.9467±0.013(4)	0.9803±0.005(3)	0.9852±0.002(4)	0.9939±0.005(2)	0.9986±0.001(2)	0.9989±0.002(2)
Medical	BMAL	0.4142±0.074(6)	0.5680±0.018(6)	0.6326±0.017(6)	0.7056±0.013(7)	0.7372±0.010(6)	0.7551±0.007(5)	0.7817±0.010(5)
	Adaptive	0.6548±0.037(2)	**0.7335±0.016**(1)	**0.7889±0.014**(1)	0.8076±0.012(2)	0.8172±0.006(3)	0.8401±0.006(2)	0.8795±0.003(3)
	BatchRank	0.3377±0.035(7)	0.4166±0.032(7)	0.5887±0.059(7)	0.7340±0.039(5)	0.8192±0.015(2)	0.8310±0.041(4)	0.8782±0.005(4)
	Random	0.6264±0.022(3)	0.7046±0.014(3)	0.7510±0.006(4)	0.7571±0.006(4)	0.7482±0.005(5)	0.7273±0.003(7)	0.7460±0.004(6)
	MGBRAML	0.6261±0.023(4)	0.6668±0.012(5)	0.7385±0.006(5)	0.7212±0.002(6)	0.7248±0.003(7)	0.7327±0.002(6)	0.7243±0.004(7)
	GBRAML_LR	**0.6552±0.043**(1)	0.7143±0.015(2)	0.7563±0.004(3)	0.7902±0.006(3)	0.8143±0.003(4)	0.8375±0.005(3)	0.9043±0.012(2)
	GBRAML_SVM	0.6176±0.057(5)	0.6887±0.024(4)	0.7674±0.010(2)	**0.8207±0.011**(1)	**0.8537±0.005**(1)	**0.8787±0.004**(1)	**0.9196±0.010**(1)
Enron	BMAL	0.4017±0.068(7)	0.4676±0.013(7)	0.4716±0.039(6)	0.4885±0.047(5)	0.4678±0.055(5)	0.4185±0.029(7)	0.4448±0.029(7)
	Adaptive	0.4832±0.009(3)	0.5121±0.003(3)	0.5326±0.003(2)	0.5434±0.004(2)	0.5462±0.002(2)	0.5438±0.002(2)	0.5721±0.002(3)
	BatchRank	0.4244±0.058(6)	0.5093±0.007(4)	0.5272±0.004(3)	0.4993±0.001(3)	0.4939±0.002(4)	0.5120±0.002(3)	0.5644±0.004(4)
	Random	0.4492±0.043(5)	0.4987±0.003(6)	0.4617±0.003(7)	0.4524±0.004(7)	0.4564±0.003(7)	0.4533±0.003(6)	0.4864±0.005(5)
	MGBRAML	0.4622±0.042(4)	0.5059±0.002(5)	0.4908±0.001(5)	0.4739±0.002(6)	0.4668±0.003(6)	0.4556±0.001(5)	0.4633±0.001(6)
	GBRAML_LR	0.4845±0.003(2)	0.5152±0.002(2)	0.5000±0.004(4)	0.4943±0.001(4)	0.4952±0.001(3)	0.5034±0.008(4)	0.5768±0.004(2)
	GBRAML_SVM	**0.5194±0.006**(1)	**0.5249±0.004**(1)	**0.5521±0.005**(1)	**0.5672±0.002**(1)	**0.5764±0.002**(1)	**0.5884±0.003**(1)	**0.6319±0.004**(1)
Slashdot	BMAL	**0.7738±0.002**(1)	0.7368±0.031(3)	0.7597±0.031(3)	0.8217±0.029(2)	**0.8462±0.001**(1)	**0.8477±0.001**(1)	**0.8593±0.002**(1)
	Adaptive	0.7189±0.003(4)	0.7246±0.003(4)	0.7414±0.001(4)	0.7395±0.049(4)	0.7997±0.001(4)	0.8302±0.004(2)	0.8513±0.001(2)
	BatchRank	0.7688±0.004(2)	**0.7857±0.008**(1)	**0.8156±0.001**(1)	**0.8325±0.001**(1)	0.8306±0.001(2)	0.8293±0.001(3)	0.8315±0.002(3)
	Random	0.5414±0.008(7)	0.6184±0.012(7)	0.5685±0.005(7)	0.5938±0.004(6)	0.6467±0.002(6)	0.6780±0.002(6)	0.7442±0.002(6)
	MGBRAML	0.6203±0.034(6)	0.6304±0.008(6)	0.6114±0.001(6)	0.5703±0.002(7)	0.5976±0.003(7)	0.6377±0.005(7)	0.6776±0.000(7)
	GBRAML_LR	0.6780±0.003(5)	0.6458±0.006(5)	0.6374±0.004(5)	0.6876±0.007(5)	0.7521±0.005(5)	0.7939±0.002(5)	0.7975±0.001(5)
	GBRAML_SVM	0.7640±0.002(3)	0.7712±0.001(2)	0.7896±0.000(2)	0.7986±0.001(3)	0.8054±0.001(3)	0.8177±0.002(4)	0.8085±0.002(4)
Languagelog	BMAL	0.0363±0.005(6)	0.0465±0.025(6)	0.0669±0.012(7)	0.0729±0.020(7)	0.0803±0.006(7)	0.0801±0.014(7)	0.0902±0.012(7)
	Adaptive	0.0386±0.005(5)	0.0524±0.005(5)	0.1102±0.007(4)	0.1489±0.001(4)	0.1937±0.002(2)	**0.2322±0.006**(1)	**0.3678±0.010**(1)
	BatchRank	0.0298±0.009(7)	0.0403±0.023(7)	0.0853±0.034(5)	0.1208±0.014(5)	0.1716±0.014(5)	0.1949±0.008(3)	0.2011±0.031(3)
	Random	0.0586±0.011(3)	0.1011±0.011(2)	**0.1726±0.008**(1)	**0.1931±0.002**(1)	**0.1980±0.005**(1)	0.1993±0.003(2)	0.1793±0.010(6)
	MGBRAML	0.0620±0.009(2)	0.0886±0.002(3)	0.1339±0.002(3)	0.1743±0.010(3)	0.1859±0.005(3)	0.1920±0.002(4)	0.1997±0.000(4)
	GBRAML_LR	**0.0819±0.012**(1)	**0.1159±0.007**(1)	0.1470±0.010(2)	0.1759±0.008(2)	0.1813±0.005(4)	0.1808±0.004(5)	0.2365±0.006(2)
	GBRAML_SVM	0.0560±0.012(4)	0.0595±0.013(4)	0.0785±0.005(6)	0.1016±0.005(6)	0.1181±0.003(6)	0.1182±0.004(6)	0.1861±0.012(5)
Bibtex	BMAL	0.2488±0.004(6)	0.2784±0.002(6)	0.3033±0.004(7)	0.3380±0.006(6)	0.3290±0.013(7)	0.3294±0.016(7)	0.3345±0.017(7)
	Adaptive	0.3046±0.002(4)	0.3454±0.002(4)	0.3998±0.001(2)	**0.4369±0.001**(1)	**0.4757±0.002**(1)	**0.4827±0.003**(1)	0.4987±0.001(3)
	BatchRank	0.2233±0.041(7)	0.2403±0.005(7)	0.3034±0.008(6)	0.3450±0.009(5)	0.3563±0.016(6)	0.3881±0.012(5)	0.4516±0.010(6)
	Random	0.3504±0.010(2)	**0.3698±0.006**(1)	0.3235±0.005(4)	0.3719±0.003(4)	0.4353±0.002(2)	0.4640±0.001(2)	0.4855±0.001(4)
	MGBRAML	**0.3579±0.002**(1)	0.3662±0.001(2)	0.3406±0.001(3)	0.3373±0.001(7)	0.3670±0.001(5)	0.3830±0.002(6)	0.4585±0.001(5)
	GBRAML_LR	0.3044±0.001(5)	0.3081±0.001(5)	0.3225±0.001(5)	0.3781±0.001(3)	0.4172±0.001(4)	0.4438±0.002(3)	0.5370±0.002(2)
	GBRAML_SVM	0.3394±0.002(3)	0.3597±0.003(3)	**0.4002±0.002**(1)	0.4136±0.001(2)	0.4223±0.001(3)	0.4371±0.001(4)	**0.5392±0.002**(1)
Total Order (Average Rank): GBRAML_SVM (2.833)<GBRAML_LR (3.071)<Adaptive (3.214)<Random (4.071)<BatchRank/MGBRAML (4.667)<BMAL (5.476)								

For AP, we observe from Fig. 5(a)–(f) that in most cases, the state-of-the-art algorithms achieve unsatisfactory performance and gain limited progress as instances are gradually selected. In contrast, the AP values deduced from GBRAML, especially for GBRAML_LR and GBRAML_SVM, are rather impressive. The underlying reason is that both GBRAML

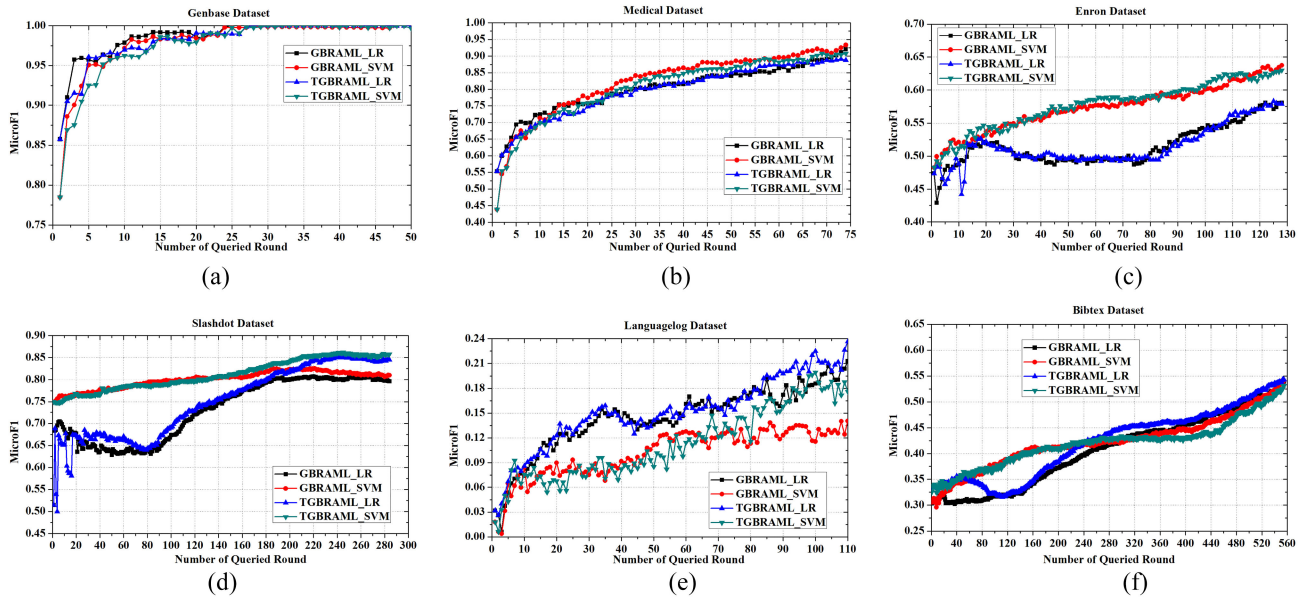

Fig. 6. Comparison of GBRAML and TGBRAML on MicroF1. (Figures best viewed in color.) (a) Genbase. (b) Medical. (c) Enron. (d) Slashdot. (e) Languagelog. (f) Bibtex.

and LLSF take a second-order learning strategy, which is close to the average cardinality of considered benchmarks. The comparable performance between GBRAML_LR and GBRAML_SVM suggests the crucial contribution of GBRAML in example-based performance improvement. The functionality of middle-level granule is more important than low-level granule, as the performance on AP between *Random* and MGBRAML is analogous and in most cases, worse than GBRAML_LR among all datasets.

Combining Figs. 4 and 5, we conclude that GBRAML is more effective in example-based metric than label-based metric. The reason is that score matrix depends heavily on the fitness label correlation, which prefers the example-based metrics.

We also compare the predictive performance on MicroF1 and AP with 5, 10, 20, 30, 40, 50, and 80% of unlabeled data used as queries in Tables V and VII, respectively. For each case, the best result is highlighted in boldface, and algorithm rankings are also provided. Comparatively, for all 42 predictive results (6 datasets $\times$ 7 observations) on MicroF1, GBRAML_LR ranks in first place at 19.05% cases (8/42), in second place at 21.4% cases (9/42), in third place at 16.67% cases (7/42), and in the second half at 42.86% cases (18/42); GBRAML_SVM ranks in first place at 30.95% cases (13/42), in second place at 16.67% cases (7/42), in third place at 16.67% cases (7/42), and in the bottom three at 35.71% cases (15/42). For all 42 predictive results (6 datasets $\times$ 7 observations) on AP, GBRAML_LR ranks in first place at 7.14% cases (3/42), in second place at 35.71% cases (15/42), in third place at 33.33% cases (14/42), and in the bottom three at 23.81%(10/42) cases; GBRAML_SVM ranks in first place at 73.81% cases (31/42), in second place at 7.14% cases (3/42), in third place at 4.76% cases (2/42), and in the second half at 14.3% cases (6/42).

Furthermore, we conduct paired t-tests at 95 significance level and present the win/tie/loss counts of GBRAML_SVM versus the other methods with respect to evaluation metric MicroF1 and AP in Tables VI and VIII, respectively. The results demonstrate that in most cases, GBRAML_SVM outperforms the compared algorithms in condition that same number of instances are selected, especially on metric AP, where the dominance gains 92.38% (194/210).

TABLE VI
WIN/TIE/LOSS COUNTS OF GBRAML_SVM VERSUS THE OTHER METHODS ON MICROF1 WITH VARIED NUMBERS OF QUERIES BASED ON PAIRED t-TESTS AT 95% SIGNIFICANCE LEVEL

Algorithms	Number of Queried Round (percentage of the unlabeled data)							
	5%	10%	20%	30%	40%	50%	80%	In All
BMAL	4/1/1	5/1/0	5/1/0	5/1/0	5/0/1	5/0/1	5/0/1	34/4/4
Adaptive	4/1/1	4/1/1	3/2/1	3/1/2	4/0/2	3/0/3	4/0/2	25/5/12
BatchRank	4/1/1	3/2/1	4/1/1	4/0/2	4/0/2	4/0/2	4/0/2	27/4/11
Random	2/3/1	2/1/3	3/2/1	4/0/2	3/1/2	4/0/2	5/1/0	23/8/11
MGBRAML	2/3/1	2/1/3	4/1/1	4/1/1	5/0/1	5/0/1	5/0/1	27/6/9
In All	16/9/5	16/6/8	19/7/4	20/3/7	21/1/8	21/0/9	23/1/6	136/27/47

2) Comparison With TGBRAML: Figs. 6 and 7 report the average performance on MicroF1 and AP across the benchmarks, respectively. In each subfigure, the x-axis denotes the round of active query and the y-axis denotes the performance (i.e., MicroF1 and AP) obtained on the unlabelled set. In what follows, we explain the comparisons of GBRAML and TGBRAML with LR (that is, GBRAML_LR plotted by black line versus TGBRAML_LR plotted by blue line) and with SVM (that is, GBRAML_SVM plotted by red line versus TGBRAML_SVM plotted by green line).

For MicroF1, the performance of TGBRAML is comparable to GBRAML in the early stage and achieves better performance since 60% instances of datasets are actively selected. Comparatively, the performance of TGBRAML_SVM fluctuates more heavily than TGBRAML_LR in the first 10% rounds, and gains more superiority than the corresponding controlled group after 50% instances are selected. This observation shows that three-way weighing schema may not be recommended if

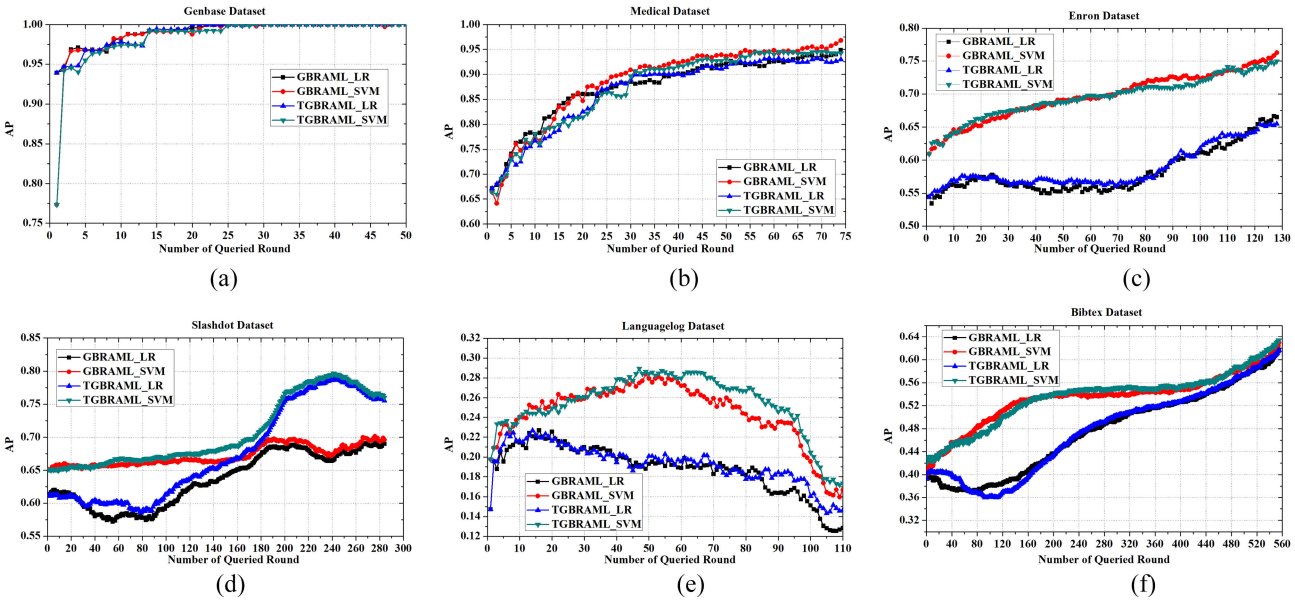

Fig. 7. Comparison of GBRAML and TGBRAML on AP. (Figures best viewed in color.) (a) Genbase. (b) Medical. (c) Enron. (d) Slashdot. (e) Languagelog. (f) Bibtex.

TABLE VII
EXPERIMENTAL RESULTS OF EACH COMPARISON ALGORITHM (MEAN±STD WITH ALGORITHM RANKING FOLLOWED AND SUMMARIZED) ON AP

Data set	Algorithms	Number of Queried Round (percentage of the unlabeled data)						
		5%	10%	20%	30%	40%	50%	80%
Genbase	BMAL	0.5122±0.039(5)	0.5819±0.047(5)	0.6590±0.005(5)	0.6880±0.033(5)	0.7107±0.031(5)	0.6947±0.063(5)	0.7166±0.032(5)
	Adaptive	0.0918±0.000(7)	0.0914±0.001(7)	0.0908±0.001(7)	0.0904±0.001(7)	0.0907±0.001(7)	0.0892±0.000(7)	0.0570±0.000(7)
	BatchRank	0.4182±0.076(6)	0.5433±0.065(6)	0.6329±0.017(6)	0.6299±0.045(6)	0.6914±0.026(6)	0.6118±0.087(6)	0.5724±0.033(6)
	Random	0.9710±0.019(2)	0.9874±0.006(2)	**0.9936±0.001**(1)	**0.9945±0.001**(1)	0.9954±0.000(3)	0.9959±0.001(3)	0.9977±0.001(3)
	MGBRAML	**0.9722±0.021**(1)	**0.9916±0.001**(1)	0.9925±0.001(2)	0.9930±0.000(3)	0.9923±0.000(4)	0.9916±0.000(4)	0.9948±0.000(4)
	GBRAML_LR	0.9579±0.015(3)	0.9681±0.002(3)	0.9899±0.006(3)	0.9934±0.002(2)	**0.9996±0.000**(1)	**1.0000±0.000**(1)	**1.0000±0.000**(1)
	GBRAML_SVM	0.9243±0.085(4)	0.9677±0.001(4)	0.9873±0.003(4)	0.9910±0.002(4)	0.9994±0.001(2)	0.9996±0.001(2)	0.9994±0.001(2)
Medical	BMAL	0.1741±0.022(5)	0.2050±0.023(6)	0.2796±0.020(6)	0.3345±0.010(6)	0.3463±0.015(6)	0.3553±0.008(6)	0.4024±0.024(5)
	Adaptive	0.0338±0.000(7)	0.0338±0.000(7)	0.0341±0.000(7)	0.0331±0.001(7)	0.0333±0.000(7)	0.0333±0.000(7)	0.0344±0.001(7)
	BatchRank	0.1735±0.035(6)	0.2372±0.014(5)	0.3027±0.017(5)	0.3488±0.018(5)	0.3793±0.015(5)	0.3785±0.010(5)	0.3233±0.013(6)
	Random	**0.7334±0.030**(1)	**0.8090±0.024**(1)	**0.8460±0.006**(1)	0.8553±0.003(3)	0.8509±0.005(3)	0.8427±0.003(4)	0.8487±0.003(4)
	MGBRAML	0.7264±0.026(2)	0.7692±0.008(3)	0.8258±0.004(4)	0.8258±0.003(4)	0.8355±0.004(4)	0.8478±0.001(3)	0.8549±0.003(3)
	GBRAML_LR	0.7184±0.035(3)	0.7744±0.010(2)	0.8425±0.013(2)	0.8649±0.007(2)	0.8843±0.003(2)	0.9004±0.003(2)	0.9400±0.005(2)
	GBRAML_SVM	0.7027±0.047(4)	0.7608±0.009(4)	0.8346±0.016(3)	**0.8822±0.009**(1)	**0.9121±0.005**(1)	**0.9380±0.030**(1)	**0.9583±0.007**(1)
Enron	BMAL	0.1170±0.007(6)	0.1432±0.006(6)	0.1751±0.007(6)	0.1778±0.009(6)	0.1872±0.011(6)	0.1877±0.007(6)	0.2161±0.009(6)
	Adaptive	0.0650±0.000(7)	0.0634±0.000(7)	0.0642±0.000(7)	0.0648±0.000(7)	0.0647±0.000(7)	0.0658±0.000(7)	0.0667±0.000(7)
	BatchRank	0.5621±0.005(2)	0.5778±0.002(2)	0.5913±0.002(2)	0.5657±0.002(2)	0.5662±0.002(2)	0.5900±0.004(2)	0.6550±0.005(3)
	Random	0.5531±0.007(5)	0.5641±0.002(4)	0.5423±0.007(5)	0.5300±0.003(5)	0.5260±0.001(5)	0.5257±0.001(4)	0.5702±0.004(4)
	MGBRAML	0.5558±0.007(4)	0.5630±0.004(5)	0.5569±0.002(4)	0.5372±0.003(4)	0.5315±0.003(4)	0.5184±0.001(5)	0.5298±0.001(5)
	GBRAML_LR	0.5591±0.005(3)	0.5673±0.005(3)	0.5623±0.002(3)	0.5546±0.003(3)	0.5544±0.003(3)	0.5752±0.007(3)	0.6614±0.004(2)
	GBRAML_SVM	**0.6356±0.008**(1)	**0.6480±0.004**(1)	**0.6733±0.004**(1)	**0.6885±0.004**(1)	**0.6956±0.002**(1)	**0.7174±0.003**(1)	**0.7555±0.005**(1)
Languagelog	BMAL	0.0319±0.002(6)	0.0388±0.004(6)	0.0513±0.007(6)	0.0649±0.004(6)	0.0713±0.010(6)	0.0768±0.003(6)	0.0875±0.008(6)
	Adaptive	0.0190±0.001(7)	0.0185±0.000(7)	0.0159±0.000(7)	0.0159±0.000(7)	0.0202±0.000(7)	0.0187±0.000(7)	0.0274±0.001(7)
	BatchRank	0.0368±0.004(5)	0.0574±0.004(5)	0.0692±0.006(5)	0.0700±0.004(5)	0.0753±0.003(5)	0.0832±0.002(5)	0.1001±0.005(5)
	Random	0.2106±0.006(3)	0.2423±0.006(2)	0.2539±0.002(2)	0.2617±0.003(2)	0.2568±0.002(2)	0.2563±0.002(2)	**0.2654±0.005**(1)
	MGBRAML	0.2017±0.009(4)	0.2169±0.006(4)	0.2330±0.002(3)	0.2564±0.003(3)	0.2558±0.002(3)	0.2554±0.002(3)	0.2487±0.000(2)
	GBRAML_LR	0.2191±0.005(2)	0.2271±0.005(3)	0.2266±0.004(4)	0.2339±0.003(4)	0.2249±0.003(4)	0.2028±0.003(4)	0.1608±0.004(4)
	GBRAML_SVM	**0.2436±0.006**(1)	**0.2556±0.005**(1)	**0.2689±0.005**(1)	**0.2908±0.006**(1)	**0.2987±0.003**(1)	**0.2856±0.004**(1)	0.1876±0.005(3)
Slashdot	BMAL	0.0778±0.021(6)	0.0822±0.009(6)	0.0909±0.010(6)	0.1182±0.007(5)	0.1414±0.003(5)	0.1424±0.002(6)	0.1690±0.019(5)
	Adaptive	0.0517±0.006(7)	0.0511±0.002(7)	0.0527±0.004(7)	0.0515±0.003(7)	0.0523±0.005(7)	0.0542±0.001(7)	0.0581±0.001(7)
	BatchRank	0.0804±0.002(5)	0.0933±0.002(5)	0.1041±0.003(5)	0.1132±0.004(6)	0.1277±0.002(6)	0.1595±0.003(5)	0.1396±0.003(6)
	Random	0.6162±0.001(3)	0.5976±0.001(3)	0.5521±0.002(4)	0.5648±0.002(3)	0.6021±0.001(3)	0.6185±0.001(3)	0.6391±0.001(3)
	MGBRAML	0.6231±0.002(2)	0.6072±0.001(2)	0.5777±0.003(3)	0.5452±0.002(4)	0.5643±0.002(4)	0.5866±0.003(4)	0.6209±0.000(4)
	GBRAML_LR	0.6146±0.002(4)	0.5894±0.003(4)	0.5824±0.001(2)	0.6026±0.003(2)	0.6352±0.002(2)	0.6741±0.003(2)	0.6879±0.003(2)
	GBRAML_SVM	**0.6572±0.001**(1)	**0.6562±0.002**(1)	**0.6600±0.000**(1)	**0.6637±0.001**(1)	**0.6632±0.001**(1)	**0.6874±0.002**(1)	**0.6958±0.002**(1)
Bibtex	BMAL	0.1725±0.013(5)	0.1945±0.005(5)	0.2296±0.004(5)	0.2493±0.008(5)	0.2426±0.011(5)	0.2292±0.007(6)	0.2106±0.015(5)
	Adaptive	0.0118±0.010(7)	0.0122±0.007(7)	0.0134±0.003(7)	0.0134±0.002(7)	0.0149±0.001(7)	0.0139±0.002(7)	0.0152±0.001(7)
	BatchRank	0.1712±0.004(6)	0.1911±0.008(6)	0.2202±0.003(6)	0.2362±0.003(6)	0.2351±0.007(6)	0.2377±0.006(5)	0.1918±0.013(6)
	Random	0.4283±0.004(3)	0.4304±0.004(3)	0.3929±0.003(3)	0.4470±0.002(2)	0.5087±0.001(2)	0.5388±0.002(2)	0.5752±0.001(3)
	MGBRAML	0.4338±0.002(2)	0.4320±0.001(2)	0.4073±0.001(2)	0.4070±0.002(4)	0.4336±0.003(4)	0.4580±0.002(4)	0.5323±0.001(4)
	GBRAML_LR	0.3778±0.001(4)	0.3714±0.001(4)	0.3928±0.002(4)	0.4449±0.002(3)	0.4882±0.001(3)	0.5132±0.001(3)	0.6133±0.003(2)
	GBRAML_SVM	**0.4429±0.003**(1)	**0.4690±0.005**(1)	**0.5247±0.002**(1)	**0.5376±0.001**(1)	**0.5389±0.002**(1)	**0.5438±0.001**(1)	**0.6244±0.003**(1)
Total Order (Average Rank): GBRAML_SVM (1.595)<GBRAML_LR (2.738)<Random (2.809)<MGBRAML (3.333)<BatchRank (4.952)<BMAL (5.571)<Adaptive (7)								

instances are rarely labeled, and is effective if the randomness of pairwise label correlation is decreased. An interesting phenomenon is that TGBRAML is most effective for datasets like "Slashdot" and "Languagelog," which does not dominates the state-of-the-art algorithms when GBRAML is applied.

TABLE VIII
WIN/TIE/LOSS COUNTS OF GBRAML_SVM VERSUS THE OTHER METHODS ON AP WITH VARIED NUMBERS OF QUERIES BASED ON PAIRED t-TESTS AT 95% SIGNIFICANCE LEVEL

Algorithms	Number of Queried Round (percentage of the unlabeled data)							
	5%	10%	20%	30%	40%	50%	80%	In All
BMAL	6/0/0	6/0/0	6/0/0	6/0/0	6/0/0	6/0/0	6/0/0	42/0/0
Adaptive	6/0/0	6/0/0	6/0/0	6/0/0	6/0/0	6/0/0	6/0/0	42/0/0
BatchRank	6/0/0	6/0/0	6/0/0	6/0/0	6/0/0	6/0/0	6/0/0	42/0/0
Random	4/2/0	4/0/2	4/1/1	5/0/1	6/0/0	6/0/0	5/0/1	34/3/5
MGBRAML	4/2/0	4/1/1	4/1/1	5/1/0	6/0/0	6/0/0	5/0/1	34/5/3
In All	26/4/0	26/1/3	26/2/2	28/1/1	30/0/0	30/0/0	28/0/2	194/8/8

For AP, the performance of TGBRAML is better than GBRAML, especially after 60% instances of datasets are actively selected. The comparative performance between TGBRAML_LR and TGBRAML_SVM is more similar than GBRAML_LR and GBRAML_SVM as active query proceeds.

Compared with Figs. 6 and 7, it is demonstrated that TGBRAML achieves better performance than GBRAML. The performance variation on AP for TGBRAML is more robust than that on MicroF1. Three-way decisions is more beneficial for the improvement of example-based criterion.

VI. CONCLUSION

This article proposes a novel batch-mode AL algorithm, GBRAML, for multilabel learning. GBRAML intends to explore the informativeness and representativeness of unlabelled instances hierarchically, which circumvents the problem of NP-hard optimization. Extensive experiments demonstrate the superiority of GBRAML. Three-way decisions is effective in measuring label correlation, and the combination of three-way decisions and GBRAML (i.e., TGBRAML) outperforms the GBRAML. In the future, we plan to extend the GBRAML for AL on large-scale multilabel datasets, which selects uncertain instances by employing shared discriminative features and evaluating hierarchical label correlation consistency based on modal-dependent deep neural network models.

REFERENCES

[1] M. L. Zhang and Z. H. Zhou, "A review on multi-label learning algorithms," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 8, pp. 1819–1837, Aug. 2014.

[2] E. Gibaja and S. Ventura, "A tutorial on multilabel learning," *ACM Comput. Surveys*, vol. 47, no. 3, pp. 1–38, Apr. 2015.

[3] S. M. Tabatabaei, S. Dick, and W. S. Xu, "Toward non-intrusive load monitoring via multi-label classification," *IEEE Trans. Smart Grid*, vol. 8, no. 1, pp. 26–40, Jan. 2017.

[4] H. Z. Fu, J. Cheng, Y. W. Xu, D. W. K. Wong, J. Liu, and X. C. Cao, "Joint optic disc and cup segmentation based on multi-label deep network and polar transformation," *IEEE Trans. Med. Imag.*, vol. 37, no. 7, pp. 1597–1605, Jul. 2018.

[5] Y. C. Wei *et al.*, "HCP: A flexible CNN framework for multi-label image classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 9, pp. 1901–1907, Sep. 2016.

[6] B. Settles, "Active learning literature survey," Dept. Comput. Sci., Univ. Wisconsin-Madison, Madison, WI, USA, Rep. 1648, 2009.

[7] Z. Wang and J. P. Ye, "Querying discriminative and representative samples for batch mode active learning," *ACM Trans. Knowl. Discov. Data*, vol. 9, no. 3, p. 17, Feb. 2015.

[8] S.-J. Huang, J. Rong, and Z.-H. Zhou, "Active learning by querying informative and representative examples," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 10, pp. 1936–1949, Oct. 2014.

[9] O. Reyes and S. Ventura, "Evolutionary strategy to perform batch-mode active learning on multi-label data," *ACM Trans. Intell. Syst. Technol.*, vol. 9, no. 4, p. 46, Feb. 2018.

[10] S. Chakraborty, V. Balasubramanian, Q. Sun, S. Panchanathan, and J. P. Ye, "Active batch selection via convex relaxations with guaranteed solution bounds," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 10, pp. 1945–1958, Oct. 2015.

[11] J. T. Yao, A. V. Vasilakos, and W. Pedrycz, "Granular computing: Perspectives and challenges," *IEEE Trans. Cybern.*, vol. 43, no. 6, pp. 1977–1989, Dec. 2013.

[12] M. Wang, Y. Lin, F. Min, and D. Liu, "Cost-sensitive active learning through statistical methods," *Inf. Sci.*, vol. 501, pp. 460–482, Oct. 2019.

[13] Y. Y. Yao, "Three-way decision and granular computing," *Int. J. Approx. Reason.*, vol. 103, pp. 107–123, Dec. 2018.

[14] Q. H. Zhang, S. H. Yang, and G. Y. Wang, "Measuring uncertainty of probabilistic rough set model from its three regions," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 47, no. 12, pp. 3299–3309, Dec. 2017.

[15] Y. J. Zhang, D. Q. Miao, Z. F. Zhang, J. F. Xu, and S. Luo, "A three-way selective ensemble model for multi-label classification," *Int. J. Approx. Reason.*, vol. 103, pp. 394–413, Dec. 2018.

[16] Y. J. Zhang, D. Q. Miao, W. Pedrycz, T. N. Zhao, J. F. Xu, and Y. Yu, "Granular structure-based incremental updating for multi-label classification," *Knowl. Based Syst.*, vol. 189, Feb. 2020, Art. no. 105066.

[17] F. Min, F.-L. Liu, L.-Y. Wen, and Z.-H. Zhang, "Tri-partition cost-sensitive active learning through kNN," *Soft Comput.*, vol. 23, no. 5, pp. 1557–1572, Mar. 2019.

[18] X. Kang, X. F. Shi, Y. N. Wu, and F. J. Ren, "Active learning with complementary sampling for instructing class-biased multi-label text emotion classification," *IEEE Trans. Affect. Comput.*, early access, Nov. 16, 2020, doi: 10.1109/TAFFC.2020.3038401.

[19] Y. F. Fu, X. Q. Zhu, and B. Li, "A survey on instance selection for active learning," *Knowl. Inf. Syst.*, vol. 35, pp. 249–283, May 2013.

[20] X. F. He, and D. Cai, "Active subspace learning," in *Proc. 12th Int. Conf. Comput. Vis.*, Kyoto, Japan, 2009, pp. 911–916.

[21] C. Persello and L. Bruzzone, "Active learning for domain adaptation in the supervised classification of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 11, pp. 4468–4483, Nov. 2012.

[22] L. H. Zhang, J. Y. Sun, T. T. Wang, Y. F. Min, and H. C. Lu, "Visual saliency detection via kernelized subspace ranking with active learning," *IEEE Trans. Image Process.*, vol. 29, pp. 2258–2270, 2020.

[23] E. A. Cherman, Y. Papanikolaou, G. Tsoumakas, and M. C. Monard, "Multi-label active learning: Key issues and a novel query strategy," *Evol. Syst.*, vol. 10, no. 1, pp. 63–78, Mar. 2019.

[24] X. Li, L. Wang, and E. Sung, "Multilabel SVM active learning for image classification," in *Proc. Int. Conf. Image Process.*, Singapore, 2004, pp. 2207–2210.

[25] B. Yang, J.-T. Sun, T. Wang, and Z. Chen, "Effective multi-label active learning for text classification," in *Proc. 15th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, 2009, pp. 917–925.

[26] X. Li and Y. H. Guo, "Active learning with multi-label SVM classification," in *Proc. 23rd Int. Joint Conf. Artif. Intell.*, Beijing, China, 2013, pp. 1479–1485.

[27] N. N. Gao, S. J. Huang, and S. C. Chen, "Multi-label active learning by model guided distribution matching," *Front. Comput. Sci.*, vol. 10, no. 5, pp. 845–855, Oct. 2016.

[28] B. Du, Z. M. Wang, L. F. Zhang, L. P. Zhang, and D. C. Tao, "Robust and discriminative labeling for multi-label active learning based on maximum correntropy criterion," *IEEE Trans. Image Process.*, vol. 26, pp. 1694–1707, 2017.

[29] S. Gharbi, M. W. Mkaouer, I. Jenhani, and M. B. Messaoud, "On the classification of software change messages using multi-label active learning," in *Proc. 34th ACM/SIGAPP Symp. App. Comput.*, Limassol, Cyprus, 2019, pp. 1760–1767.

[30] S. C. H. Hoi, R. Jin, and M. R. Lyu, "Batch mode active learning with applications to text categorization and image retrieval," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 9, pp. 1233–1248, Sep. 2009.

[31] S. Chakraborty, V. Balasubramanian, and S. Panchanathan, "Optimal batch selection for active learning in multi-label classification," in *Proc. 19th ACM Int. Conf. Multimedia*, Scottsdale, AZ, USA, 2011, pp. 1413–1416.

[32] B. Zhang, Y. Wang, and F. Chen, "Multilabel image classification via high-order label correlation driven active learning," *IEEE Trans. Image Process.*, vol. 23, pp. 1430–1441, 2014.

[33] X. G. You, R. X. Wang, and D. C. Tao, "Diverse expected gradient active learning for relative attributes," *IEEE Trans. Image Process.*, vol. 23, pp. 3203–3217, 2014.

[34] B. Demir, C. Persello, and L. Bruzzone, "Batch-mode active-learning methods for the interactive classification of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 3, pp. 1014–1031, Mar. 2011.

[35] S. Patra and L. Bruzzone, "A batch-mode active learning technique based on multiple uncertainty for SVM classifier," *IEEE Geosci. Remote Sens. Lett.*, vol. 9, no. 3, pp. 497–501, May 2012.
[36] T. N. C. Cardoso, R. M. Silva, S. Canuto, M. M. Moro, and M. A. Gonçalves, "Ranked batch-mode active learning," *Inf. Sci.*, vol. 379, pp. 313–337, Feb. 2017.
[37] S. Kee, E. D. Castillo, and G. Runger, "Query-by-committee improvement with diversity and density in batch active learning," *Inf. Sci.*, vols. 454–455, pp. 401–418, Jul. 2018.
[38] J. Yuan, X. X. Hou, Y. Q. Xiao, D. Cao, W. L. Guan, and L. Q. Nie, "Multi-criteria active deep learning for image classification," *Knowl. Based Syst.*, vol. 172, pp. 86–94, May 2019.
[39] J. Wu, A. Q. Guo, V. S. Sheng, P. P. Zhao, and Z. M. Cui, "An active learning approach for multi-label image classification with sample noise," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 32, no. 3, Mar. 2018, Art. no. 1850005.
[40] Z. C. Qiu, D. J. Miller, and G. Kesidis, "A maximum entropy framework for semisupervised and active learning with unknown and label-scarce classes," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 4, pp. 917–933, Apr. 2017.
[41] Y.-F. Yan and S.-J. Huang, "Cost-effective active learning for hierarchical multi-label classification," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Stockholm, Sweden, 2018, pp. 2962–2968.
[42] M.-L. Zhang and L. Wu, "LIFT: Multi-label learning with label-specific features," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 1, pp. 107–120, Jan. 2015.
[43] J. Huang, G. R. Li, Q. M. Huang, and X. D. Wu, "Learning label-specific features and class-dependent labels for multi-label classification," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 12, pp. 3309–3323, Dec. 2016.
[44] M. Oakes, R. Gaizauskas, H. Fowkes, A. Jonsson, V. Wan, and M. Beaulieu, "A method based on the chi-square test for document classification," in *Proc. 24th ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, New Orleans, LA, USA, 2001, pp. 440–441,
[45] X. Y. Jia, W. W. Li, J. Y. Liu, and Y. Zhang, "Label distribution learning by exploiting label correlations," in *Proc. 23rd Int. Joint Conf. Artif. Intell.*, Beijing, China, 2018, pp. 3310–3317.
[46] K. Pearson and A. Lee, "On the laws of inheritance in man: I. Inheritance of physical characters," *Biometrika*, vol. 2, no. 4, pp. 357–462, 1902.
[47] D. C. Liang, W. Pedrycz, and D. Liu, "Determining three-way decisions with decision-theoretic rough sets using a relative value approach," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 47, no. 8, pp. 1785–1799, Aug. 2017.

Yuanjian Zhang received the Ph.D. degree in computer science and technology from Tongji University, Shanghai, China, in 2020.

He is currently a Postdoctoral Fellow jointly trained by the Fudan University, Shanghai, and China UnionPay Research Institute of Electornic Payment, Shanghai. His major interests include multilabel learning, machine learning, and big data analysis.

Tianna Zhao received the master's degree in probability theory and mathematical statistics from Hebei Normal University, Shijiazhuang, China, in 2018. She is currently pursuing the Ph.D. degree with the Department of Computer Science and Technology, Tongji University, Shanghai, China.

Her major interests include multilabel learning and label distribution learning.

Duoqian Miao received the Ph.D. degree in pattern recognition and intelligent system from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 1997.

He is a Professor with Tongji University, Shanghai, China. His research includes pattern recognition and big data analysis. He has published more than 100 academic papers in international journals, including IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, TC, and TFS.

Prof. Miao is a Fellow of International Rough Set Society and Chinese Association for Artificial Intelligence.

Witold Pedrycz (Life Fellow, IEEE) received Ph.D. degree from the Silesian University of Technology, Gliwice, Poland, in 1980.

He is a Professor and a Canada Research Chair (CRC-Computational Intelligence) with the Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB, Canada. He is with the Systems Research Institute of the Polish Academy of Sciences, Warsaw, Poland. Since 2019, he has been an expert in high-level visiting professor program with Tongji University. In 2009, he was elected as a Foreign Member of the Polish Academy Sciences. He has published numerous papers in this area. He has authored 14 research monographs covering various aspects of computational intelligence and software engineering. His main research directions include computational intelligence, fuzzy modeling and granular computing, knowledge discovery and data mining, fuzzy control, pattern recognition, knowledge-based neural networks, relational computing, and software engineering.

Prof. Pedrycz is intensively involved in editorial activities. He is the Editor-in-Chief of Information Sciences. He currently serves on the Advisory Board of IEEE TRANSACTIONS ON FUZZY SYSTEMS. He is a member of a number of editorial boards of other international journals. He has been a member of numerous program committees of IEEE conferences in the area of fuzzy sets and neurocomputing.