Active Learning by Querying Informative and Representative Examples

Sheng-Jun Huang, Rong Jin, *Member, IEEE*, and Zhi-Hua Zhou, *Fellow, IEEE*

Abstract—Active learning reduces the labeling cost by iteratively selecting the most valuable data to query their labels. It has attracted a lot of interests given the abundance of unlabeled data and the high cost of labeling. Most active learning approaches select either informative or representative unlabeled instances to query their labels, which could significantly limit their performance. Although several active learning algorithms were proposed to combine the two query selection criteria, they are usually ad hoc in finding unlabeled instances that are both informative and representative. We address this limitation by developing a principled approach, termed QUIRE, based on the min-max view of active learning. The proposed approach provides a systematic way for measuring and combining the informativeness and representativeness of an unlabeled instance. Further, by incorporating the correlation among labels, we extend the QUIRE approach to multi-label learning by actively querying instance-label pairs. Extensive experimental results show that the proposed QUIRE approach outperforms several state-of-the-art active learning approaches in both single-label and multi-label learning.

Index Terms—Active learning, learning with unlabeled data, multi-label learning, informativeness, representativeness

1 Introduction

In many real-world problems, unlabeled data are often abundant whereas labeled data are scarce. Label acquisition is usually expensive due to the involvement of human experts, and thus, it is important to train an accurate prediction model by a small number of labeled instances. Active learning addresses this challenge by querying only the most valuable instances for their class assignments [37].

The key component of an active learning algorithm lies in the design of an appropriate criterion for selecting the most valuable instances for querying, a problem that is often referred to as *query selection*. Two types of query selection criteria, i.e., *informativeness* and *representativeness*, are widely used by active learning algorithms. Informativeness measures the ability of an instance in reducing the uncertainty of a statistical model, whereas representativeness measures whether an instance well represents the overall input patterns of unlabeled data [37]. Most active learning algorithms deploy only one of the two criteria for query selection, which could significantly limit their performance. In particular, approaches favoring informative instances usually do not exploit the structure of unlabeled data, leading to serious sample bias and consequently undesirable performance; approaches favoring representative instances may have to query a relatively large number of instances before the optimal decision boundary is found. Although several active learning methods [11], [27], [47] were developed to find the instances that are both informative and representative, they are mostly ad hoc in measuring the informativeness and representativeness of an instance, leading to suboptimal performance.

In this paper, we propose a novel approach for active learning by QUerying Informative and Representative Examples (*QUIRE* for short). QUIRE is based on the min-max view of active learning [19], which provides a systematic way for measuring and combining the two query selection criteria. More specifically, QUIRE measures both the informativeness and representativeness of an instance by its prediction uncertainty: the informativeness of an instance $\mathbf{x}$ is measured by its prediction uncertainty according to the labeled data, whereas the representativeness of $\mathbf{x}$ is measured by its prediction uncertainty according to the unlabeled data. By applying similar measures to both criteria, QUIRE is effective in identifying queries that are both informative and representative, which is verified by our empirical study.

The second contribution of this work is to extend the QUIRE approach to multi-label learning [53], a setting that is much less studied in active learning. Unlike single-label learning where one instance is assumed to be associated with only one label, in multi-label learning, instances can be assigned to multiple labels simultaneously. Many real-world problems can be cast into multi-label learning, including image annotation [4] and text classification [45]. Because one needs to decide, for each label, its relevance to an instance, the labeling cost is much higher for multi-label learning than that for single-label learning, and therefore, active query mechanisms are highly desirable for multi-label learning. We further improve the QUIRE algorithm by incorporating into the query selection process the label correlation, which is known to be crucial for multi-label learning [21], [33].

- *S.-J. Huang and Z.-H. Zhou are with the National Key Laboratory of Novel Software Technology, Nanjing University, Nanjing 210023, China. E-mail: {huangsj, zhouzh}@lamda.nju.edu.cn.*
- *R. Jin is with the Department of Computer Science and Engineering, Michigan State University, East Lansing, MI 48824. E-mail: rongjin@cse.msu.edu.*

Manuscript received 3 Sept. 2013; revised 30 Dec. 2013; accepted 16 Feb. 2014. Date of publication 23 Feb. 2014; date of current version 10 Sept. 2014. Recommended for acceptance by J. Ye.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TPAMI.2014.2307881

0162-8828 © 2014 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Huazhong University of Science and Technology. Downloaded on November 02,2020 at 09:30:26 UTC from IEEE Xplore. Restrictions apply.

The rest of this paper is organized as follows: Section 2 reviews some related work; Section 3 presents our proposed approach under the single-label setting, which is then extended to multi-label learning in Section 4; experimental results are reported in Section 5; Section 6 concludes this work with future issues.

2 Related Work

Querying the most informative instances is probably the most popular approach for active learning. Exemplar approaches include query-by-committee [9], [16], [38], uncertainty sampling [2], [25], [26], [41], expected error reduction based sampling [35] and mutual information based sampling [17], [18]. The main weakness of these approaches lies in the fact that they are unable to exploit the abundance of unlabeled data and the query selection is solely determined by a small number of labeled examples, making it prone to sample bias.

Another school of active learning is to select the instances that are most representative to the unlabeled data. Most approaches in this group aim to exploit the cluster structure of unlabeled data [8], [10], [30], usually by a clustering method. The main weakness of them lies in the fact that their performance heavily depends on the quality of clustering results [10]. Optimal experimental design methods also try to query representative examples [15], [50], but usually ignore the information of the queried labels.

Several active learning algorithms tried to combine the informativeness measure with the representativeness measure for finding the optimal query instances. A representative sampling algorithm [47] is to exploit the cluster information of unlabeled instances as well as the classification margin. One limitation of this approach is that clustering is only performed on the instances within the classification margin, leaving the unlabeled instances outside the margin unexploited. In [11], Donmez et al. extended the active learning approach in [30] by dynamically balancing the uncertainty and the density of instances for query selection. This approach is ad hoc in combining the measure of informativeness and representativeness for query selection, leading to suboptimal performance. Recently, Wang and Ye [46] derived an empirical upper bound for active learning risk, and by minimizing this upper bound, a batch model active learning method was proposed to select instances that are discriminative and with similar distribution to the unlabeled data. However, because the number of instances selected at each iteration is usually quite small, the distribution estimated on the very limited amount of data could be less accurate.

Our work is based on the min-max view of active learning, which was first proposed in the study of batch mode active learning [19]. Unlike [19] which measures the representativeness of an instance by its similarity to the remaining unlabeled instances, our proposed measure of representativeness takes into account the cluster structure of unlabeled instances as well as the class assignments of the labeled examples, leading to a better selection of instances for query.

Compared to single-label learning, active learning under multi-label setting is much less studied. Multi-label learning, where one instance can be simultaneously associated with multiple labels, has attracted many research interests during the past few years [31], [40], [48], [53]. The task of multi-label learning is to learn a mapping from the feature space to the label space, which consists of the power set of all labels and could be extremely large. To handle such a challenging task, it has been shown that it is important to exploit the correlation between labels [21], [44], [51].

Most active learning algorithms decompose a multi-label task into a set of binary classification problems. For example, in [5], [13] and [39], uncertainty are first measured for each label, and then combined to form the uncertainty measure for individual instances. In [29], one SVM classifier is trained for each label, and the instance leading to the maximum reduction of expected loss is selected. Similarly, in [49], by introducing an extra regression model to predict the number of class labels that will be assigned to each instance, the expected loss reduction based on independently trained SVMs is used as the selection criterion. This work is further improved in [23] by the introduction of an auxiliary learner. Recently, Li and Guo proposed to measure the informativeness of an instance by combining the label cardinality inconsistency and the separation margin with a tradeoff parameter [28].

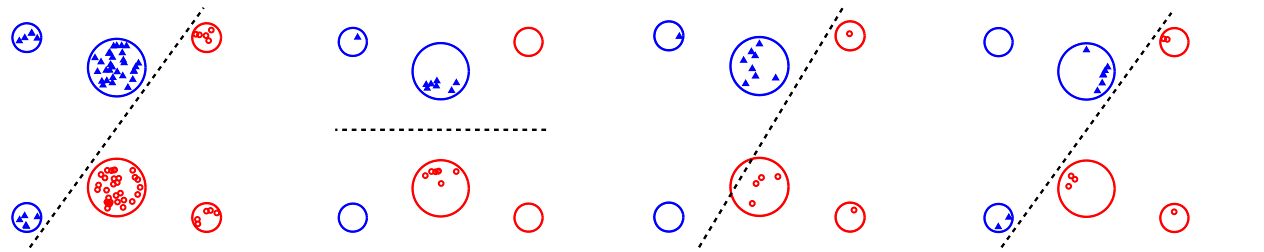
While most active learning algorithms are designed to query *all* the label assignments of the selected instances, Qi et al. proposed a two-dimensional approach in [32] that queries *instance-label pairs*; in other words, it selects *one* label c and an instance $\mathbf{x}$, and queries the oracle if $\mathbf{x}$ should be assigned to label c. Huang and Zhou [22] follows this setting, and selects instance-label pairs based on a label ranking model. Since the strategy of querying instance-label pairs is shown to be more effective than querying all the label assignments [32], we adopt the strategy in this study.

The main limitation of existing multi-label active learning approaches lies in the fact that they are restricted to selecting the most informative instances. In addition, most of them treat multiple labels independently, ignoring the correlations among labels, which has shown to be crucial to multi-label learning [51], [53]. We address these limitations by combining label correlation with the measures of representativeness and informativeness for query selection.

3 QUIRE for Single-Label Learning

To illustrate the importance of querying instances that are both informative and representative for active learning, we first perform a empirical study on a synthetic data set. Fig. 1a shows the synthetic data set for binary classification, where each class is represented by a different legend. We examine three different active learning algorithms by allowing them to sequentially select 15 data points. Figs. 1b and 1c show the data points selected by an approach favoring informative instances (i.e., [41]) and by an approach favoring representative instances (i.e., [10]), respectively. As indicated by Fig. 1b, due to the sample bias, the approach preferring informative instances tends to choose the data points close to the horizontal line, leading to incorrect decision boundaries. On the other hand, as indicated by Fig. 1c, the approach preferring representative instances is able to identify the approximately correct decision boundary but

Authorized licensed use limited to: Huazhong University of Science and Technology. Downloaded on November 02,2020 at 09:30:26 UTC from IEEE Xplore. Restrictions apply.



(a) data distribution (b) instances queried based on informativeness (c) instances queried based on representativeness (d) instances queried by our method

Fig. 1. An illustrative example for selecting informative and representative instances.

with a slow convergence because it does not favor the informative instances. Fig. 1d shows the data points selected by our proposed approach that favors data points that are both informative and representative. It is clear that our proposed algorithm is more efficient in finding the accurate decision boundary than the other two approaches.

We denote by $\mathcal{D} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_{n_l}, y_{n_l}), \mathbf{x}_{n_l+1}, \ldots, \mathbf{x}_n\}$ the training data set that consists of n_l labeled instances and $n_u = n - n_l$ unlabeled instances, where each instance $\mathbf{x}_i = [x_{i1}, x_{i2}, \ldots, x_{id}]^\top$ is a vector of d dimension and $y_i \in \{-1, +1\}$ is the class label of $\mathbf{x}_i$. Active learning selects one instance $\mathbf{x}_s$ from the pool of unlabeled data to query its label. The goal is to learn an accurate model by labeling as few unlabeled instances as possible. For convenience, we divide the data set $\mathcal{D}$ into three parts: the labeled data $\mathcal{D}_l$, the currently selected instance $\mathbf{x}_s$, and the rest of the unlabeled data $\mathcal{D}_u$. We also use $\mathcal{D}_a = \mathcal{D}_u \cup \{\mathbf{x}_s\}$ to represent all the unlabeled instances. We use $\mathbf{y} = [\mathbf{y}_l, y_s, \mathbf{y}_u]$ for the label assignment for the entire data set, where $\mathbf{y}_l$, y_s and $\mathbf{y}_u$ are the labels assigned to $\mathcal{D}_l$, $\mathbf{x}_s$ and $\mathcal{D}_u$, respectively. Finally, we denote by $\mathbf{y}_a = [y_s, \mathbf{y}_u]$ the label assignment for all the unlabeled instances.

3.1 The Framework

In order to motivate the proposed approach for active learning, we first re-examine the margin-based active learning approach from the viewpoint of min-max by following the discussion in [19]. Let f^* be a classification model trained by the labeled examples, i.e.,

$$f^* = \arg\min_{f \in \mathcal{H}} \frac{\lambda}{2}|f|^2_{\mathcal{H}} + \sum_{i=1}^{n_l} \ell(y_i, f(\mathbf{x}_i)), \tag{1}$$

where $\mathcal{H}$ is a reproducing kernel Hilbert space endowed with kernel function $\kappa(\cdot,\cdot) : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$. $\ell(z)$ is the loss function. Given classifier f^*, the margin-based approach chooses the unlabeled instance closest to the decision boundary, i.e.,

$$s^* = \arg\min_{n_l < s \le n} |f^*(\mathbf{x}_s)|. \tag{2}$$

Proposition 1 connects the margin based query selection with the min-max formulation of active learning.

Proposition 1. *The criterion in Eq. (2) can be rewritten as*

$$s^* = \arg\min_{n_1 < s \le n} \mathcal{L}(\mathcal{D}_l, \mathbf{x}_s), \tag{3}$$

where

$$\mathcal{L}(\mathcal{D}_l, \mathbf{x}_s) = \max_{y_s = \pm 1} \min_{f \in \mathcal{H}} \frac{\lambda}{2}|f|^2_{\mathcal{H}} + \sum_{i=1}^{n_l} \ell(y_i, f(\mathbf{x}_i)) + \ell(y_s, f(\mathbf{x}_s)). \tag{4}$$

Proof. Denote by $\mathcal{J}(f)$ the object function, i.e.,

$$\mathcal{J}(f) = \frac{\lambda}{2}|f|^2_{\mathcal{H}} + \sum_{i=1}^{n_l} \ell(y_i, f(\mathbf{x}_i)),$$

we have

$$\begin{aligned} s^* &= \arg\min_{n_l < s \le n} |f^*(\mathbf{x}_s)| \\ &= \arg\min_{n_l < s \le n} \min_{f \in \mathcal{H}; f: \mathcal{J}(f) \le \mathcal{J}(f^*)} |f(x_s)| \\ &= \arg\min_{n_l < s \le n} \min_{f \in \mathcal{H};} |f(x_s)| + C\mathcal{J}(f) \\ &= \arg\min_{n_l < s \le n} \max_{y_s = \pm 1} \min_{f \in \mathcal{H}} \ell(y_s, f(\mathbf{x}_s)) + C\mathcal{J}(f) \\ &= \arg\min_{n_l < s \le n} \max_{y_s = \pm 1} \min_{f \in \mathcal{H}} C\left(\frac{\lambda}{2}|f|^2_{\mathcal{H}} + \sum_{i=1}^{n_l} \ell(y_i, f(\mathbf{x}_i))\right) \\ &\quad + \ell(y_s, f(\mathbf{x}_s)). \end{aligned}$$

Let $C = 1$, we have $s^* = \arg\min_{n_1 < s \le n} \mathcal{L}(\mathcal{D}_l, \mathbf{x}_s)$. □

Further we can write Eq. (3) in a minimax form

$$s^* = \arg\min_{n_l < s \le n} \max_{y_s = \pm 1} A(\mathcal{D}_l, \mathbf{x}_s),$$

where

$$A(\mathcal{D}_l, \mathbf{x}_s) = \min_{f \in \mathcal{H}} \frac{\lambda}{2}|f|^2_{\mathcal{H}} + \sum_{i=1}^{n_l} \ell(y_i, f(\mathbf{x}_i)) + \ell(y_s, f(\mathbf{x}_s)).$$

In this min-max view of active learning, it guarantees that the selected instance $\mathbf{x}_s$ will lead to a small value for the objective function regardless of its class label y_s. In order to select queries that are both informative and representative, we extend the evaluation function $\mathcal{L}(\mathcal{D}_l, \mathbf{x}_s)$ to include all the unlabeled data. Hypothetically, if we know the class assignment $\mathbf{y}_u$ for the unselected unlabeled instances in $\mathcal{D}_u$, the evaluation function can be modified as

$$\mathcal{L}(\mathcal{D}_l, \mathcal{D}_u, \mathbf{y}_u, \mathbf{x}_s) = \max_{y_s = \pm 1} \min_{f \in \mathcal{H}} \frac{\lambda}{2}|f|^2_{\mathcal{H}} + \sum_{i=1}^{n} \ell(y_i, f(\mathbf{x}_i)). \tag{5}$$

Authorized licensed use limited to: Huazhong University of Science and Technology. Downloaded on November 02,2020 at 09:30:26 UTC from IEEE Xplore. Restrictions apply.

The problem is that the class assignment $\mathbf{y}_u$ is unknown. According to the manifold assumption [3], we expect that a correct solution for $\mathbf{y}_u$ should result in a small value of $\mathcal{L}(\mathcal{D}_l, \mathcal{D}_u, \mathbf{y}_u, \mathbf{x}_s)$. We therefore approximate the solution for $\mathbf{y}_u$ by minimizing $\mathcal{L}(\mathcal{D}_l, \mathcal{D}_u, \mathbf{y}_u, \mathbf{x}_s)$, which leads to the following evaluation function for query selection:

$$\begin{aligned}\widehat{\mathcal{L}}(\mathcal{D}_l, \mathcal{D}_u, \mathbf{x}_s) &= \min_{\mathbf{y}_u \in \{\pm 1\}^{n_u - 1}} \mathcal{L}(\mathcal{D}_l, \mathcal{D}_u, \mathbf{y}_u, \mathbf{x}_s) \\ &= \min_{\mathbf{y}_u \in \{\pm 1\}^{n_u - 1}} \max_{y_s = \pm 1} \min_{f \in \mathcal{H}} \frac{\lambda}{2} |f|^2_{\mathcal{H}} + \sum_{i=1}^{n} \ell(y_i, f(\mathbf{x}_i)). \end{aligned} \tag{6}$$

As a result, we will find instance $\mathbf{x}_s$ that minimizes the evaluation function $\widehat{\mathcal{L}}(\mathcal{D}_l, \mathcal{D}_u, \mathbf{x}_s)$. In the next section, we will discuss how to efficiently compute the evaluation function $\widehat{\mathcal{L}}(\mathcal{D}_l, \mathcal{D}_u, \mathbf{x}_s)$.

3.2 The Solution

For the computational simplicity, for the rest of this work, we choose a quadratic loss function, i.e., $\ell(y, \widehat{y}) = (y - \widehat{y})^2/2$.[1] It is straightforward to show

$$\min_{f \in \mathcal{H}} \frac{\lambda}{2} |f|^2_{\mathcal{H}} + \frac{1}{2} \sum_{i=1}^{n} (y_i - f(\mathbf{x}_i))^2 = \frac{1}{2} \mathbf{y}^\top L \mathbf{y},$$

where $L = (K + \lambda I)^{-1}$ and $K = [\kappa(\mathbf{x}_i, \mathbf{x}_j)]_{n \times n}$ is the kernel matrix of size $n \times n$. Thus, the evaluation function $\widehat{\mathcal{L}}(\mathcal{D}_l, \mathcal{D}_u, \mathbf{x}_s)$ is simplified as

$$\widehat{\mathcal{L}}(\mathcal{D}_l, \mathcal{D}_u, \mathbf{x}_s) = \min_{\mathbf{y}_u \in \{-1,+1\}^{n_u - 1}} \max_{y_s \in \{-1,+1\}} \mathbf{y}^\top L \mathbf{y}. \tag{7}$$

Our goal is to efficiently compute the above quantity for each unlabeled instance. For the convenience of presentation, we refer to by subscript u the rows/columns in a matrix M for the unlabeled instances in $\mathcal{D}_u$, by subscript l the rows/columns in M for labeled instances in $\mathcal{D}_l$, and by subscript s the row/column in M for the selected instance. We also refer to by subscript a the rows/columns in M for all the unlabeled instances (i.e., $\mathcal{D}_u \cup \{\mathbf{x}_s\}$). Using these conventions, we rewrite the objective $\mathbf{y}^\top L \mathbf{y}$ as

$$\begin{aligned}\mathbf{y}^\top L \mathbf{y} &= \mathbf{y}_l L_{l,l} \mathbf{y}_l + L_{s,s} + \mathbf{y}_u^T L_{u,u} \mathbf{y}_u \\ &\quad + 2\mathbf{y}_u^\top (L_{u,l} \mathbf{y}_l + L_{u,s} y_s) + 2 y_s \mathbf{y}_l^\top L_{l,s}. \end{aligned} \tag{8}$$

Note that since the above objective function is concave (linear) in y_s and convex (quadratic) in $\mathbf{y}_u$, we can switch the maximization of $\mathbf{y}_u$ with the minimization of y_s in (7). By relaxing $\mathbf{y}_u$ to continuous variables, the solution to $\min_{\mathbf{y}_u} \mathbf{y}^\top L \mathbf{y}$ is given by

$$\widehat{\mathbf{y}}_u = -L_{u,u}^{-1} (L_{u,l} \mathbf{y}_l + L_{u,s} y_s), \tag{9}$$

leading to the following expression for the evaluation function $\widehat{\mathcal{L}}(\mathcal{D}_l, \mathcal{D}_u, \mathbf{x}_s)$:

$$\begin{aligned}&\widehat{\mathcal{L}}(\mathcal{D}_l, \mathcal{D}_u, \mathbf{x}_s) \\ &\quad = L_{s,s} + \mathbf{y}_l^T L_{l,l} \mathbf{y}_l + \max_{y_s = \pm 1} \{ 2 y_s L_{s,l} \mathbf{y}_l \\ &\qquad - (L_{u,l} \mathbf{y}_l + L_{u,s} y_s)^\top L_{u,u}^{-1} (L_{u,l} \mathbf{y}_l + L_{u,s} y_s) \} \\ &\quad \propto L_{s,s} - \frac{\det(L_{a,a})}{L_{s,s}} + 2 \Big| \Big(L_{s,l} - L_{s,u} L_{u,u}^{-1} L_{u,l} \Big) \mathbf{y}_l \Big|, \end{aligned} \tag{10}$$

where the last step follows the relation

$$\det\left(\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \right) = \det(A_{22}) \det\big(A_{11} - A_{12} A_{22}^{-1} A_{21}\big).$$

Here we do not require the prediction of unlabeled data, i.e., $\mathbf{y}_u$ to be accurate because it is used only as an intermediate quantity to facilitate the measure of representativeness for unlabeled instances. It is also worth to note that although the evaluation function (10) is derived under the binary classification setting, it can be easily extended to multi-class learning with the one-versus-rest scheme. Formally, assume that there are m classes, then the evaluation function can be modified as:

$$\begin{aligned}\widehat{\mathcal{L}}(\mathcal{D}_l, \mathcal{D}_u, \mathbf{x}_s) = L_{s,s} + \max_{j = 1 \cdots m} \{ &\mathbf{y}_l^{j\top} L_{l,l} \mathbf{y}_l^j + 2 L_{s,l} \mathbf{y}_l^j \\ &- \big(L_{u,l} \mathbf{y}_l^j + L_{u,s}\big)^\top L_{u,u}^{-1} \big(L_{u,l} \mathbf{y}_l^j + L_{u,s}\big) \},\end{aligned}$$

where $\mathbf{y}_l^j$ is the labels of labeled data on the jth class.

Remark. The evaluation function $\widehat{\mathcal{L}}(\mathcal{D}_l, \mathcal{D}_u, \mathbf{x}_s)$ essentially consists of two components: $L_{s,s} - \det(L_{a,a})/L_{s,s}$ and $|(L_{s,l} - L_{s,u} L_{u,u}^{-1} L_{u,l}) \mathbf{y}_l|$. Minimizing the first component is equivalent to minimizing $L_{s,s}$ because $L_{a,a}$ is independent from the selected instance $\mathbf{x}_s$. Since $L = (K + \lambda I)^{-1}$, we have

$$\begin{aligned} L_{s,s} &= \left[K_{s,s} - (K_{s,l}, K_{s,u}) \begin{pmatrix} K_{l,l} & K_{l,u} \\ K_{u,l} & K_{u,u} \end{pmatrix} \begin{pmatrix} K_{l,s} \\ K_{u,s} \end{pmatrix} \right]^{-1} \\ &\approx \frac{1}{K_{s,s}} \left[1 + \frac{(K_{s,l}, K_{s,u})}{K_{s,s}} \begin{pmatrix} K_{l,l} & K_{l,u} \\ K_{u,l} & K_{u,u} \end{pmatrix} \begin{pmatrix} K_{l,s} \\ K_{u,s} \end{pmatrix} \right]. \end{aligned}$$

Therefore, to choose an instance with small $L_{s,s}$, we select the instance with large self-similarity $K_{s,s}$. When self-similarity $K_{s,s}$ is a constant, this term will have no effect for query selection.

To analyze the effect of the second component, we approximate it as:

$$\begin{aligned} &2 \Big| \Big(L_{s,l} - L_{s,u} L_{u,u}^{-1} L_{u,l} \Big) \mathbf{y}_l \Big| \\ &\quad \approx 2 \big| L_{s,l} \mathbf{y}_l \big| + 2 \Big| L_{s,u} L_{u,u}^{-1} L_{u,l} \mathbf{y}_l \Big| \\ &\quad \approx 2 |L_{s,l} \mathbf{y}_l| + 2 |L_{s,u} \widehat{\mathbf{y}}_u|. \end{aligned} \tag{11}$$

The first term in the above approximation measures the confidence in predicting $\mathbf{x}_s$ using only labeled data, which corresponds to the *informativeness* of $\mathbf{x}_s$. The second term measures the prediction confidence using only the predicted labels of the unlabeled data, which can be viewed as the measure of *representativeness*. This is because when $\mathbf{x}_s$ is a representative instance, it is expected to share a

1. Although quadratic loss may not be ideal for classification, it does yield competitive classification results when compared to the other loss functions such as hinge loss [34].

Authorized licensed use limited to: Huazhong University of Science and Technology. Downloaded on November 02,2020 at 09:30:26 UTC from IEEE Xplore. Restrictions apply.

large similarity with many of the unlabeled instances. As a result, the prediction for $\mathbf{x}_s$ by the unlabeled data in $\mathcal{D}_u$ is decided by the average of their assigned class labels $\widehat{\mathbf{y}}_u$. If we assume that the classes are evenly distributed over the unlabeled data, we should expect a low confidence in predicting the class label for $\mathbf{x}_s$ by unlabeled data. Note that unlike the existing work that measures the representativeness only by the cluster structure of unlabeled data, the proposed measure of representativeness depends on $\widehat{\mathbf{y}}_u$, which essentially combines the cluster structure of unlabeled data with the class assignments of labeled data. Given the high dimensional data, there could be many possible cluster structures that are consistent with the unlabeled data and it is unclear which one is consistent with the target classification problem. It is therefore critical to take into account the label information when exploiting the cluster structure of unlabeled data. Here note the approximation in Eq. (11) is derived only for analysis, our algorithm is based on the minimax principle instead of the combination of two criteria.

3.3 Efficient Algorithm

Computing the evaluation function $\widehat{\mathcal{L}}(\mathcal{D}_l, \mathcal{D}_u, \mathbf{x}_s)$ in Eq. (10) requires computing $L_{u,u}^{-1}$ for every unlabeled instance $\mathbf{x}_s$, leading to high computational cost when the number of unlabeled instances is very large. The theorem below allows us to improve the computation efficiency dramatically.

Theorem 2. *Let*

$$L_{a,a}^{-1} = \begin{pmatrix} L_{s,s} & L_{s,u} \\ L_{u,s} & L_{u,u} \end{pmatrix}^{-1} = \begin{pmatrix} a & -\mathbf{b}^\top \\ -\mathbf{b} & D \end{pmatrix}.$$

We have $L_{u,u}^{-1} = D - \frac{1}{a}\mathbf{b}\mathbf{b}^\top$.

Proof. Using the matrix inversion lemma, we have

$$\begin{aligned} L_{a,a}^{-1} &= \begin{pmatrix} L_{s,s} & L_{s,u} \\ L_{u,s} & L_{u,u} \end{pmatrix}^{-1} = \begin{pmatrix} a & -\mathbf{b}^\top \\ -\mathbf{b} & D \end{pmatrix} \\ &= \begin{pmatrix} C_1^{-1} & -\frac{1}{L_{s,s}} L_{u,s}^\top C_2^{-1} \\ -\frac{1}{L_{s,s}} C_2^{-1} L_{u,s} & C_2^{-1} \end{pmatrix}, \end{aligned}$$

$$\text{where} \quad C_1 = L_{s,s} - L_{u,s}^\top L_{u,u}^{-1} L_{u,s},$$

$$C_2 = L_{u,u} - \frac{1}{L_{s,s}} L_{u,s} L_{u,s}^\top.$$

With the equation above, we can express a, $\mathbf{b}$ and D in terms of L as follows:

$$\begin{aligned} \frac{1}{a} &= C_1 = L_{s,s} - L_{u,s}^T L_{u,u}^{-1} L_{u,s}, \\ D &= C_2^{-1} = \left(L_{u,u} - \frac{1}{L_{s,s}} L_{u,s} L_{u,s}^\top \right)^{-1} \\ &= L_{u,u}^{-1} + L_{u,u}^{-1} L_{u,s} \left(L_{s,s} - L_{u,s}^T L_{u,u}^{-1} L_{u,s} \right)^{-1} L_{u,s}^T L_{u,u}^{-1} \\ &= L_{u,u}^{-1} + a L_{u,u}^{-1} L_{u,s} L_{u,s}^T L_{u,u}^{-1}, \\ \mathbf{b} &= \frac{1}{L_{s,s}} C_2^{-1} L_{u,s} = a L_{u,u}^{-1} L_{u,s}. \end{aligned}$$

We complete the proof by combining the above relationships. □

As indicated by Theorem 2, we only need to compute $L_{a,a}^{-1}$ once since $L_{a,a}$ is independent from the selected instance $\mathbf{x}_s$. For each $\mathbf{x}_s$, its $L_{u,u}^{-1}$ can be computed directly from $L_{a,a}^{-1}$. The following proposition allows us to simplify the computation for $L_{a,a}^{-1}$.

Proposition 3. $L_{a,a}^{-1} = (\lambda I_a + K_{a,a}) - K_{a,l}(\lambda I_l + K_{l,l})^{-1} K_{l,a}$.

Proposition 3 follows directly from the inverse of a block matrix. As indicated by Proposition 3, we only need to compute $(\lambda I + K_{l,l})^{-1}$. Given that the number of labeled examples is relatively small compared to the size of unlabeled data, the computation of $L_{a,a}^{-1}$ is in general efficient. Excluding the time for computing the kernel matrix, the computational complexity of our algorithm is just $O(n_u)$. The pseudo-code of QUIRE is summarized in Algorithm 1.

Algorithm 1 The QUIRE Algorithm

Input:
  $\mathcal{D}$: a data set of n instances
Initialize:
  $\mathcal{D}_l = \varnothing$; $n_l = 0$ % no labeled data is available at the very beginning
  $\mathcal{D}_u = \mathcal{D}$; $n_u = n$ % the pool of unlabeled data
Calculate K
repeat
  Calculate $L_{a,a}^{-1}$ using Proposition 3 and $\det(L_{a,a})$
  for $s = 1$ **to** n_u **do**
    Calculate L_{uu}^{-1} according to Theorem 2
    Calculate $\widehat{\mathcal{L}}(\mathcal{D}_l, \mathcal{D}_u, \mathbf{x}_s)$ using Eq. 10
  end for
  Select the $\mathbf{x}_{s^*}$ with the smallest $\widehat{\mathcal{L}}(\mathcal{D}_l, \mathcal{D}_u, \mathbf{x}_{s^*})$
  Query the label y_{s^*} for the selected instance $\mathbf{x}_{s^*}$
  $\mathcal{D}_l = \mathcal{D}_l \cup (\mathbf{x}_{s^*}, y_{s^*})$; $\mathcal{D}_u = \mathcal{D}_u \setminus \mathbf{x}_{s^*}$
until the number of queries or the required accuracy is reached

4 QUIRE FOR MULTI-LABEL LEARNING

In this section, we extend QUIRE to multi-label learning. The most common active learning approach for multi-label learning is to solicit *all* the label assignments for each selected instance. The alterative approach is to choose one label c for each selected instance $\mathbf{x}$, and query the oracle if $\mathbf{x}$ is assigned to c, an approach that is often referred to as *instance-label pair* queries. In [32], the authors show that querying instance-label pairs is more effective because acquiring all the label assignments for the selected instances suffers from high cost. The observation is particularly true when the number of labels is large as human experts can hardly identify all relevant labels for a given instance, but can easily decide whether or not a label is relevant to the selected instance. As a result, we adopt the paradigm of querying instance-label pairs for multi-label active learning.

Let m be the number of labels, and let the label assignment of each instance $\mathbf{x}_i$ be denoted by a label vector $\mathbf{y}_i = [y_{i1}, y_{i2}, \ldots, y_{im}]^\top$, where $y_{ik} = 1$ if instance $\mathbf{x}_i$ has the kth label, and $y_{ik} = -1$ otherwise. Note that the key quantity in

Authorized licensed use limited to: Huazhong University of Science and Technology. Downloaded on November 02,2020 at 09:30:26 UTC from IEEE Xplore. Restrictions apply.

the QUIRE algorithm presented in Section 3 is the matrix L. Following the same path, to extend QUIRE algorithm to multi-label setting, we will also define an appropriate matrix L.

We first consider the simple case of active learning that does not take into account the label correlation. By learning one classifier for each label independently, the objective function of the multi-label learning task with quadratic loss can be formalized as:

$$\min_{f_k \in \mathcal{H}} \lambda \sum_{k=1}^{m} |f_k|_{\mathcal{H}}^2 + \sum_{i=1}^{n} \sum_{k=1}^{m} \left(f_k(\mathbf{x}_i) - y_{i,k}\right)^2, \tag{12}$$

where f_k is the classification model for the kth label. Let $Y = [y_{ik}]_{n \times m}$ be the ground-truth label matrix, which is partially known, and $F = [f_k(\mathbf{x}_i)]_{n \times m} = (\mathbf{f}_1, \ldots, \mathbf{f}_m)$ be the prediction matrix, where $\mathbf{f}_k$ is the predictions of all instance for the kth label. The optimization problem in Eq. (12) can be rewritten as:

$$\min_{F \in \mathbb{R}^{n \times m}} \lambda \mathrm{tr}\left(F^\top K^{-1} F\right) + |F - Y|_2^2, \tag{13}$$

where $\mathrm{tr}(\cdot)$ computes the trace of a matrix, and K is the kernel matrix.

As stated before, label correlation is critical to multi-label learning. Particularly, under the active learning setting, the information embedded in an unknown label may be inferred from some correlated labels that have been queried, avoiding the cost of querying from the oracle. Next, we introduce the label correlation into Eq. (13). Let $R \in \mathbb{R}_+^{m \times m}$ be the label correlation matrix. A straightforward approach to take into account the label correlation is to modify Eq. (13) as

$$\min_{F \in \mathbb{R}^{n \times m}} \lambda \mathrm{tr}\left(R^{-1} F^\top K^{-1} F\right) + |F - Y|_2^2. \tag{14}$$

By introducing the function $vec(\cdot)$ to convert a matrix into a vector, the solution of F in the above optimization problem is given by

$$vec(F) = [\lambda(R^{-1} \otimes K^{-1}) + I]^{-1} vec(Y), \tag{15}$$

and accordingly, the optimal value of Eq. (14) is

$$vec(Y)^\top (I - [\lambda(R^{-1} \otimes K^{-1}) + I]^{-1}) vec(Y), \tag{16}$$

where $\otimes$ is the kronecker product between matrices, and I is the identity matrix of size $nm \times nm$. To define matrix L as that for the single-label case, we write Eq. (16) as $vec(Y)^\top L vec(Y)$, such that it has the same form of Eq. (8) in Section 3, and define L as:

$$\begin{aligned} L &= I - [\lambda(R^{-1} \otimes K^{-1}) + I]^{-1} \\ &= I - [(R \otimes K)^{-1}(\lambda I + (R \otimes K))]^{-1} \\ &= \lambda[(R \otimes K) + \lambda I]^{-1}, \end{aligned}$$

where the last step follows the equation $(I + AB)^{-1} = I - A(I + BA)^{-1} B$. It is noteworthy that L encodes both the correlation between different instances and the dependence among difference labels, and is the basis for our proposed algorithm.

As we note at the beginning of this section, our goal is to query the most informative and representative instance-label pairs. Similar to Section 3, we refer to all labeled, unlabeled, and selected instance-label pairs (i.e., rows/columns in $nm \times nm$ matrix) by subscripts l, u and s, respectively. Following the same analysis as in Section 3, we have the solution for Y_u, i.e., the unlabeled instance-label pairs in Y as

$$vec(Y_u) = -L_{u,u}^{-1}(L_{u,l} vec(Y_l) + L_{u,s} Y_s). \tag{17}$$

Thus, similar to Eq. (10) in Section 3, for any instance-label pair $(\mathbf{x}_a, y_{a,b})$, its evaluation function can be obtained as:

$$\begin{aligned} \widehat{\mathcal{L}}(\mathbf{x}_a, y_{a,b}) = {} & L_{s,s} + vec(Y_l)^T L_{l,l} vec(Y_l) \\ & + \max_{y_{a,b}} \{2 y_{a,b} L_{s,l} vec(Y_l) - (L_{u,l} vec(Y_l) \\ & + L_{u,s} y_{a,b})^T L_{u,u}^{-1} (L_{u,l} vec(Y_l) + L_{u,s} y_{a,b})\}. \end{aligned} \tag{18}$$

Using the evaluation function $\widehat{\mathcal{L}}(\mathbf{x}_a, y_{a,b})$, at each iteration of active learning, we calculate the value of $\widehat{\mathcal{L}}(\mathbf{x}_a, y_{a,b})$ for every unlabeled instance-label pair $(\mathbf{x}_a, y_{a,b})$, and choose the one $(\mathbf{x}_{a^*}, y_{a^*,b^*})$ with minimal value to query, i.e.,

$$(a^*, b^*) = \arg\min_{a,b} \widehat{\mathcal{L}}(\mathbf{x}_a, y_{a,b}). \tag{19}$$

It is straightforward to verify that all the tricks developed in Section 3 for speeding up computation can be directly applied to the multi-label version algorithm.

5 Experiments

We first present the experiments for single-label tasks, followed by the experiments for multi-label learning.

5.1 Study on Single-Label Data

5.1.1 Settings

Under the single-label setting, we compare QUIRE with the following five baseline approaches:

- RANDOM: randomly selecting query instances.
- MARGIN: margin-based active learning [41], an approach that prefers informative instances.
- CLUSTER: hierarchical-clustering-based active learning [10], an approach that prefers representative instances.
- IDE: active learning that selects informative and diverse examples [19].
- DUAL: a dual strategy that exploits both informativeness and representativeness.

Note that IDE is designed for batch mode active learning, we turn it into active learning with selection of a single instance by setting the parameter $k = 1$.

Twelve data sets are used in our study and their characteristics are summarized in Table 1. *Digit1* and *g241n* are benchmark data for semi-supervised learning [7]. *Austria*, *isolet*, *titato*, *vechicle*, and *wdbc* are UCI data sets [1]. *Letter*

Authorized licensed use limited to: Huazhong University of Science and Technology. Downloaded on November 02,2020 at 09:30:26 UTC from IEEE Xplore. Restrictions apply.

TABLE 1
Data Set Information, Including the Number of Instances and the Number of Features

Data	# ins.	# feature	Data	# ins.	# feature
austra	690	14	*wdbc*	569	30
digit1	1500	241	*letterEvsF*	1543	16
g241n	1500	241	*letterIvsJ*	1502	16
isolet	600	617	*letterMvsN*	1575	16
titato	958	9	*letterDvsP*	1608	16
vehicle	435	18	*letterUvsV*	1577	16

is a multi-class data set [1], from which we select five pairs of letters that are relatively difficult to distinguish, i.e., *D versus P*, *E versus F*, *I versus J*, *M versus N*, *U versus V*, and construct a binary class data set for each pair. Each data set is randomly divided into two parts of equal size, with one part as the test data and the other part as the unlabeled data for active learning. We assume that no labeled data is available at the very beginning of active learning. For MARGIN, IDE and DUAL, instances are randomly selected when no classification model is available, which only takes place at the beginning. In each iteration, an unlabeled instance is first selected to solicit its label and the classification model is then retrained. We evaluate the classification model by its performance on the holdout test data. Both classification accuracy and area under ROC curve (AUC) are used for evaluation metrics. For every data set, we run the experiment 10 times, each with a random partition of the data set. In all the experiments, a RBF kernel is used and the parameter λ is set to be 1. LibSVM [6] is used to train a SVM classifier for all approaches in comparison.

5.1.2 Comparison with State-of-the-Art Methods

Fig. 2 shows the classification accuracy of different active learning approaches with varied numbers of queries. Table 2 shows the AUC values, with 5, 10, 20, 30, 40, 50 and 80 percent of unlabeled data used as queries. For each case, the best result and its comparable performances are highlighted in boldface based on paired t-tests at 95 percent significance level. Table 3 presents the win/tie/loss counts of QUIRE versus the other methods based on the same test.

First, we observe that the RANDOM approach tends to yield decent performance when the number of queries is very small. But, as the number of queries increases, this simple approach loses its edge and often is not as effective as the other active learning approaches. MARGIN, the most commonly used approach for active learning, is not performing well at the beginning of the learning stage. As the number of queries increases, we observe that MARGIN catches up with the other approaches and yields decent performance. This phenomenon can be attributed to the fact that with only a few training examples, the learned decision boundary tends to be inaccurate, and as a result, the unlabeled instances closest to the decision boundary may not be the most informative ones. The performance of CLUSTER is mixed. It works well on some data sets, but performs poorly on the others. We attribute the inconsistency of CLUSTER to the fact that cluster structure of unlabeled data may not be consistent with the target classification model.

The behavior of IDE is similar to that of CLUSTER in that it achieves good performance on certain data sets and fails on the others. DUAL does not yield good performance on

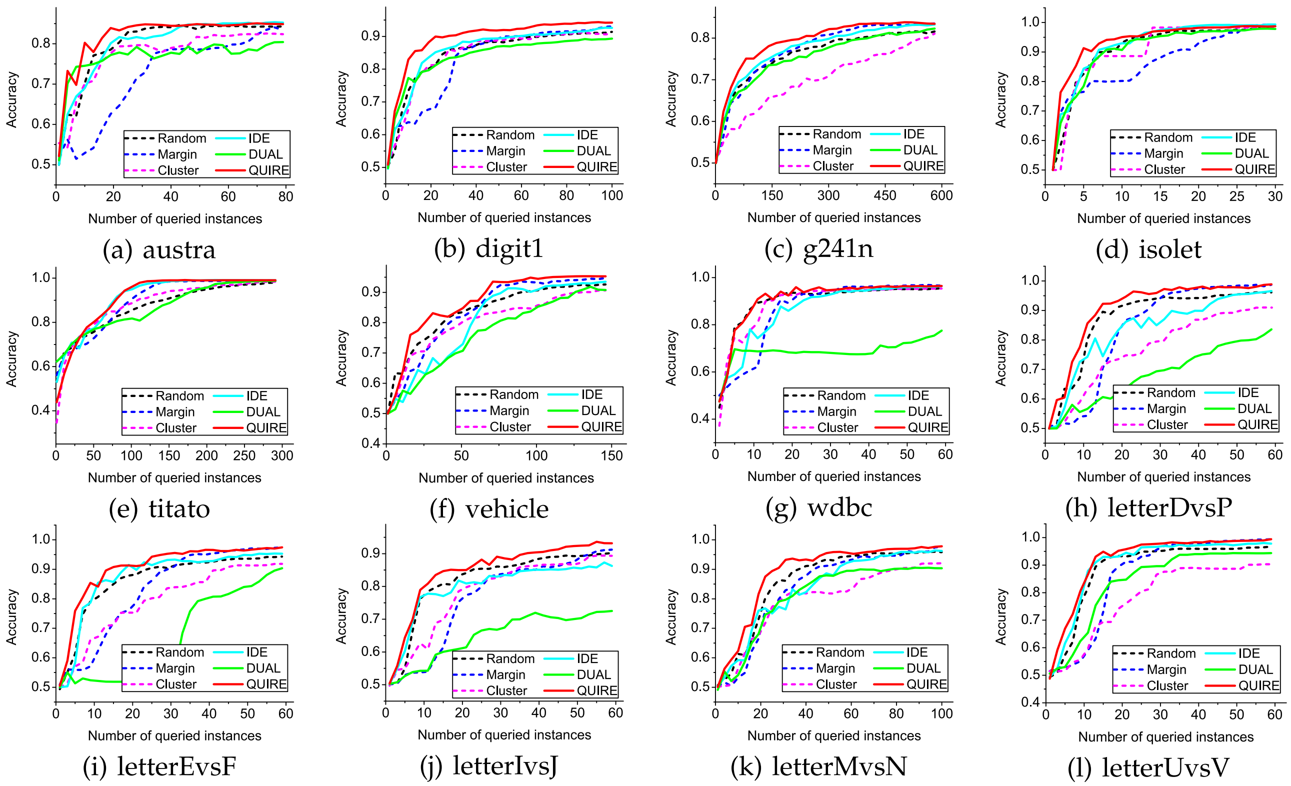


Fig. 2. Comparison on classification accuracy.

Authorized licensed use limited to: Huazhong University of Science and Technology. Downloaded on November 02,2020 at 09:30:26 UTC from IEEE Xplore. Restrictions apply.

TABLE 2
Comparison on AUC Values (Mean ± std)

Data	Algorithms	Number of queries (percentage of the unlabeled data)						
		5%	10%	20%	30%	40%	50%	80%
austra	RANDOM	**.868±.027**	**.894±.022**	**.897±.023**	**.901±.022**	**.909±.015**	**.909±.012**	**.917±.011**
	MARGIN	.751±.137	**.838±.119**	**.885±.043**	**.909±.010**	**.911±.012**	**.914±.009**	**.915±.008**
	CLUSTER	**.877±.045**	**.888±.029**	**.894±.015**	.896±.015	.903±.014	**.907±.015**	**.913±.011**
	IDE	**.858±.101**	**.885±.058**	**.902±.012**	**.912±.008**	**.913±.009**	**.914±.007**	**.916±.007**
	DUAL	**.866±.037**	**.878±.036**	.875±.018	.876±.016	.879±.013	.881±.013	.904±.008
	QUIRE	**.887±.014**	**.901±.010**	**.906±.016**	**.912±.009**	**.914±.009**	**.915±.007**	**.916±.007**
digit1	RANDOM	.945±.009	.969±.006	.979±.005	.984±.003	.985±.003	.988±.003	.991±.002
	MARGIN	.941±.028	.972±.009	**.989±.002**	**.992±.002**	**.992±.002**	**.992±.002**	**.992±.002**
	CLUSTER	.938±.035	.952±.018	.963±.019	.974±.011	.985±.002	.988±.003	**.992±.002**
	IDE	.954±.011	.973±.007	.987±.002	**.991±.002**	**.992±.002**	**.992±.002**	**.992±.002**
	DUAL	.929±.014	.953±.009	.975±.004	.982±.005	.985±.003	.987±.003	.991±.002
	QUIRE	**.976±.006**	**.986±.003**	**.990±.002**	.992±.002	**.992±.002**	**.992±.002**	**.992±.002**
g241n	RANDOM	**.713±.040**	.769±.021	.822±.018	.854±.016	.873±.015	.886±.012	.906±.014
	MARGIN	.700±.057	.751±.048	.830±.022	.864±.019	**.896±.012**	**.911±.008**	**.918±.008**
	CLUSTER	.720±.038	.770±.024	.815±.018	.835±.021	.860±.022	.880±.013	.909±.009
	IDE	.727±.030	.786±.029	**.840±.017**	.866±.016	.883±.013	.899±.011	.916±.010
	DUAL	.722±.040	.751±.019	.822±.011	.838±.022	.865±.016	.881±.012	.912±.007
	QUIRE	**.757±.035**	**.825±.019**	**.857±.020**	**.884±.013**	**.900±.009**	**.912±.006**	**.920±.009**
isolet	RANDOM	**.995±.006**	**.998±.002**	**.999±.001**	**1.00±.000**	**1.00±.000**	**1.00±.000**	**1.00±.000**
	MARGIN	**.965±.052**	**.999±.001**	**1.00±.000**	**1.00±.000**	**1.00±.000**	**1.00±.000**	**1.00±.000**
	CLUSTER	**.998±.002**	**.999±.002**	**1.00±.000**	1.00±.000	**1.00±.000**	**1.00±.000**	**1.00±.000**
	IDE	**.998±.003**	**.999±.002**	**.999±.001**	**1.00±.001**	**1.00±.000**	**1.00±.000**	**1.00±.000**
	DUAL	**.993±.008**	**.999±.001**	.999±.001	**1.00±.000**	**1.00±.001**	**1.00±.000**	**1.00±.000**
	QUIRE	**.997±.002**	**.999±.001**	.999±.001	1.00±.000	**1.00±.001**	**1.00±.000**	**1.00±.000**
titato	RANDOM	**.762±.033**	.861±.031	.954±.023	.979±.011	.991±.007	.997±.004	1.00±.000
	MARGIN	.645±.096	.753±.078	.946±.043	.998±.001	**1.00±.000**	**1.00±.000**	1.00±.000
	CLUSTER	**.717±.087**	.806±.054	.908±.031	.971±.021	.989±.010	.997±.003	**1.00±.000**
	IDE	**.735±.040**	**.906±.029**	**.996±.003**	**.999±.001**	**1.00±.001**	**1.00±.000**	**1.00±.000**
	DUAL	.708±.069	.782±.064	.900±.027	.981±.012	.995±.006	.999±.001	**1.00±.000**
	QUIRE	**.736±.037**	.861±.025	.991±.004	**.999±.001**	**1.00±.000**	**1.00±.000**	**1.00±.000**
vehicle	RANDOM	**.818±.064**	.864±.039	.925±.032	.949±.026	.968±.016	.975±.013	.989±.006
	MARGIN	.693±.078	.828±.077	**.883±.105**	**.981±.014**	**.993±.005**	**.993±.005**	**.992±.005**
	CLUSTER	**.771±.088**	.845±.056	.927±.022	.955±.018	.973±.010	.978±.011	**.992±.006**
	IDE	**.731±.141**	**.849±.106**	.878±.093	.957±.037	.977±.010	.985±.009	**.991±.006**
	DUAL	.680±.074	.706±.114	.817±.061	.875±.035	.908±.035	.947±.035	.980±.016
	QUIRE	**.750±.137**	**.912±.024**	**.956±.025**	**.985±.007**	.989±.006	.991±.005	**.992±.005**
wdbc	RANDOM	**.984±.006**	.986±.005	.990±.004	.991±.004	.991±.004	.991±.004	**.993±.003**
	MARGIN	**.967±.038**	**.990±.002**	**.993±.003**	**.993±.003**	**.993±.003**	**.993±.003**	**.993±.003**
	CLUSTER	.981±.007	**.987±.004**	.991±.003	.992±.003	**.992±.003**	**.993±.003**	**.993±.003**
	IDE	**.983±.006**	**.984±.008**	.990±.004	.992±.003	.993±.003	.993±.003	**.993±.003**
	DUAL	.955±.025	.964±.016	.972±.015	**.988±.009**	**.992±.003**	.992±.003	.992±.004
	QUIRE	**.985±.006**	**.990±.004**	**.993±.003**	**.993±.003**	**.993±.003**	**.993±.003**	**.993±.003**
letterDvsP	RANDOM	.990±.004	.995±.002	.997±.002	.998±.001	.998±.001	.998±.001	**.999±.001**
	MARGIN	.994±.005	**.999±.001**	**.999±.000**	**.999±.001**	**.999±.001**	**.999±.001**	**.999±.001**
	CLUSTER	.988±.008	.995±.004	.997±.002	.998±.001	.999±.001	**.999±.001**	**.999±.001**
	IDE	.992±.006	.997±.002	.998±.001	.999±.001	.999±.001	.999±.001	.999±.001
	DUAL	.978±.005	.986±.001	.988±.004	.990±.004	.996±.001	.998±.001	.999±.001
	QUIRE	**.998±.001**	.999±.001	**.999±.001**	**.999±.001**	**.999±.001**	**.999±.001**	**.999±.001**
letterEvsF	RANDOM	**.977±.020**	.988±.009	.994±.002	.997±.002	.998±.001	.999±.001	1.00±.000
	MARGIN	**.987±.008**	**.999±.001**	**1.00±.000**	**1.00±.000**	**1.00±.000**	**1.00±.000**	**1.00±.000**
	CLUSTER	**.975±.016**	.991±.003	**.997±.004**	.999±.001	**1.00±.000**	**1.00±.000**	1.00±.000
	IDE	.977±.014	.995±.003	.999±.000	.999±.000	.999±.000	1.00±.000	**1.00±.000**
	DUAL	.976±.011	.993±.003	.996±.002	.996±.002	.996±.002	.998±.001	**1.00±.000**
	QUIRE	**.988±.009**	**.999±.000**	**1.00±.000**	**1.00±.000**	**1.00±.000**	**1.00±.000**	**1.00±.000**
letterIvsJ	RANDOM	**.943±.025**	**.966±.017**	.980±.004	.983±.005	.985±.005	.987±.004	**.990±.004**
	MARGIN	.882±.096	**.960±.027**	**.986±.005**	**.989±.006**	**.991±.004**	**.991±.004**	**.991±.004**
	CLUSTER	**.952±.022**	**.961±.017**	.976±.008	**.985±.007**	.987±.006	**.989±.005**	**.991±.004**
	IDE	**.934±.030**	**.969±.011**	.979±.006	.980±.006	.982±.008	.985±.005	**.990±.004**
	DUAL	.819±.120	.897±.058	.934±.030	.954±.017	.959±.014	.953±.015	.988±.004
	QUIRE	**.951±.023**	**.963±.013**	.976±.011	**.989±.010**	**.991±.004**	**.991±.004**	**.991±.004**
letterMvsN	RANDOM	.977±.010	.992±.002	.994±.003	.996±.002	.997±.001	.997±.001	**.998±.001**
	MARGIN	**.964±.040**	**.991±.014**	**.999±.000**	**.999±.000**	**.999±.000**	**.999±.000**	**.999±.000**
	CLUSTER	.971±.017	.986±.009	.994±.003	.997±.002	.998±.001	.998±.001	**.999±.000**
	IDE	.969±.017	.988±.007	.997±.002	.998±.001	.998±.001	.998±.001	**.999±.000**
	DUAL	.950±.025	.972±.011	.974±.007	.980±.008	.983±.007	.983±.007	.998±.001
	QUIRE	**.986±.007**	**.996±.003**	**.998±.001**	**.999±.000**	**.999±.000**	**.999±.000**	**.999±.000**
letterUvsV	RANDOM	.992±.005	.996±.004	.998±.001	.999±.000	1.00±.000	1.00±.000	**1.00±.000**
	MARGIN	**.998±.002**	**1.00±.000**	**1.00±.000**	**1.00±.000**	**1.00±.000**	**1.00±.000**	**1.00±.000**
	CLUSTER	.990±.008	**.996±.009**	1.00±.000	1.00±.000	1.00±.000	1.00±.000	**1.00±.000**
	IDE	.995±.004	**.999±.001**	1.00±.000	1.00±.000	1.00±.000	1.00±.000	**1.00±.000**
	DUAL	.983±.014	.986±.008	.990±.008	.991±.008	.993±.007	.995±.005	.999±.000
	QUIRE	**.999±.001**	**1.00±.000**	**1.00±.000**	**1.00±.000**	**1.00±.000**	**1.00±.000**	**1.00±.000**

The best performance and its comparable performances based on paired t-tests at 95 percent significance level are highlighted in boldface.

most data sets although we have tried our best efforts to tune the related parameters.

Finally, we observe that for most cases, the QUIRE approach is able to outperform the baseline methods significantly, as indicated by Fig. 2, Tables 2 and 3. We attribute the success of QUIRE to the principle of choosing instances that are both informative and representative, and the specially designed computational framework that

Authorized licensed use limited to: Huazhong University of Science and Technology. Downloaded on November 02,2020 at 09:30:26 UTC from IEEE Xplore. Restrictions apply.

TABLE 3
Win/Tie/Loss Counts of QUIRE versus the Other Methods with Varied Numbers of Queries Based on Paired t-Tests at 95 Percent Significance Level

Algorithms	Number of queries (percentage of the unlabeled data)							In All
	5%	10%	20%	30%	40%	50%	80%	
RANDOM	4/8/0	8/4/0	9/3/0	9/2/1	10/2/0	10/2/0	6/6/0	56/27/1
MARGIN	6/6/0	4/7/1	2/8/2	2/8/2	0/11/1	0/11/1	1/11/0	15/62/7
CLUSTER	6/6/0	7/5/0	8/4/0	11/1/0	9/3/0	6/6/0	3/9/0	50/34/0
IDE	6/6/0	6/5/1	6/5/1	8/4/0	8/4/0	8/4/0	2/10/0	44/38/2
DUAL	8/4/0	10/2/0	11/1/0	10/2/0	10/2/0	11/1/0	9/3/0	69/15/0
In All	30/30/0	35/23/2	36/21/3	40/17/3	37/22/1	35/24/1	21/39/0	234/176/10

appropriately measures and combines the informativeness and representativeness.

5.1.3 Comparison on Computational Cost

We report the average CPU time (in seconds) of each query for the compared approaches in Table 4. All the experiments are performed with MATLAB 7.6 on a 3.00 GHZ Intel(R) Core(TM)2 DUO PC running Windows 7 with 4 GB main memory. It is not surprising to observe that Margin, Cluster, and IDE are the most efficient due to the simplicity of the criteria used for selecting the informative instances. The proposed algorithm QUIRE is significantly more efficient than DUAL, because of the techniques introduced to speedup the computation. We finally note that the high computational cost of QUIRE is due to the complicated criterion we adopt for instance selection, which leads to significant advantages in classification accuracy as shown in the last section.

5.2 Study on Multi-Label Data

5.2.1 Settings

Under the multi-label setting, we compare QUIRE with five multi-label active learning approaches:

- RANDOM: randomly selects instance-label pairs.
- 2DAL: selects instance-label pairs that lead to the maximum reduction of expected error [32].
- MML: selects instances with the mean max loss to query its label [29].
- MMC: selects instances that lead to the maximum loss reduction with the largest confidence [49].
- ADAPTIVE: considers both the max-margin prediction uncertainty and the label cardinality inconsistency when selecting query instances [28].

Experiments are performed on 18 data sets, most of which are available at MULAN project.[2] *Emotions* [42] consists of 593 songs. The task is to predict the music emotions of songs. *Enron* is a subset of the Enron email corpus [24], including about 1,700 emails, where each e-mail is represented as a 1,001-dimensional feature vector. *Image* is a data set for natural scene image classification, and contains 2,000 images [52]. *Medical* is a data set of clinical text for medical classification. *Scene* contains 2,407 images with SIX possible labels: beach, sunset, fall foliage, field, mountain and urban. *Reuters* is a data set for text categorization. It is a processed version of [36] with the method introduced in [54]. *Yeast* is a data set for predicting the gene functional classes of the Yeast *Saccha-romyces cerevisiae*, we use the version preprocessed by [12], which contains 2417 genes. *Yahoo* consists of 11 independent data sets, i.e., *Arts*, *Business*, *Computers*, *Education*, *Entertainment*, *Health*, *Recreation*, *Reference*, *Science*, *Social*, and *Society*. They are collected from "yahoo.com" domain [45] for webpage categorization. Each of the 11 data sets contains 5,000 documents. Twenty to 45 percent of the documents have more than one labels.

For each data set, we randomly divide it into two parts with equal size, one as test set and the other one as the unlabeled pool for active selection. The random data partition is repeated for 10 times, and average results over the 10 repeats are reported. At the very beginning of active learning, 5 percent of the unlabeled instances are randomly sampled as initial labeled data. At each iteration of active learning, QUIRE, Random and 2DAL query one instance-label pair, while the other approaches query the entire label vector for an instance, which is equivalent to m instance-label pairs. After every $2 \times m$ instance-label pairs are queried, a new classification model will be trained on the labeled data and its performance will be evaluated on the holdout test data. We stop the querying process when all the instances are fully labeled or the number of queried instance-label pairs reaches the maximum value which is set to be $20{,}000$ in our experiments.

F1-score is used to evaluate the performances of the approaches in comparison. F1-score combines precision

TABLE 4
Average CPU Time (in Seconds) of Each Query for Compared Methods

Data	Algorithms				
	Margin	Cluster	IDE	DUAL	QUIRE
austra	0.0173	0.0072	0.0265	2.0109	0.1880
digit1	0.2018	0.0109	0.0435	9.3486	3.3787
g241n	0.3955	0.0198	0.0725	6.6166	3.3816
isolet	0.0686	0.0059	0.0284	7.9308	0.1445
titato	0.0310	0.0085	0.0335	1.8330	0.8326
vehicle	0.0057	0.0048	0.0176	0.1845	0.0535
wdbc	0.0070	0.0053	0.0224	0.5171	0.1313
DvsP	0.0311	0.0131	0.0405	5.1526	3.7448
EvsF	0.0331	0.0120	0.0395	1.1038	4.2273
IvsJ	0.0470	0.0135	0.0424	1.6074	3.6689
MvsN	0.0417	0.0121	0.0442	4.5766	3.5365
UvsV	0.0275	0.0118	0.0415	4.7951	4.6030
Average	0.0756	0.0104	0.0377	3.8064	2.3242

2. http://mulan.sourceforge.net/datasets.html.

Authorized licensed use limited to: Huazhong University of Science and Technology. Downloaded on November 02,2020 at 09:30:26 UTC from IEEE Xplore. Restrictions apply.

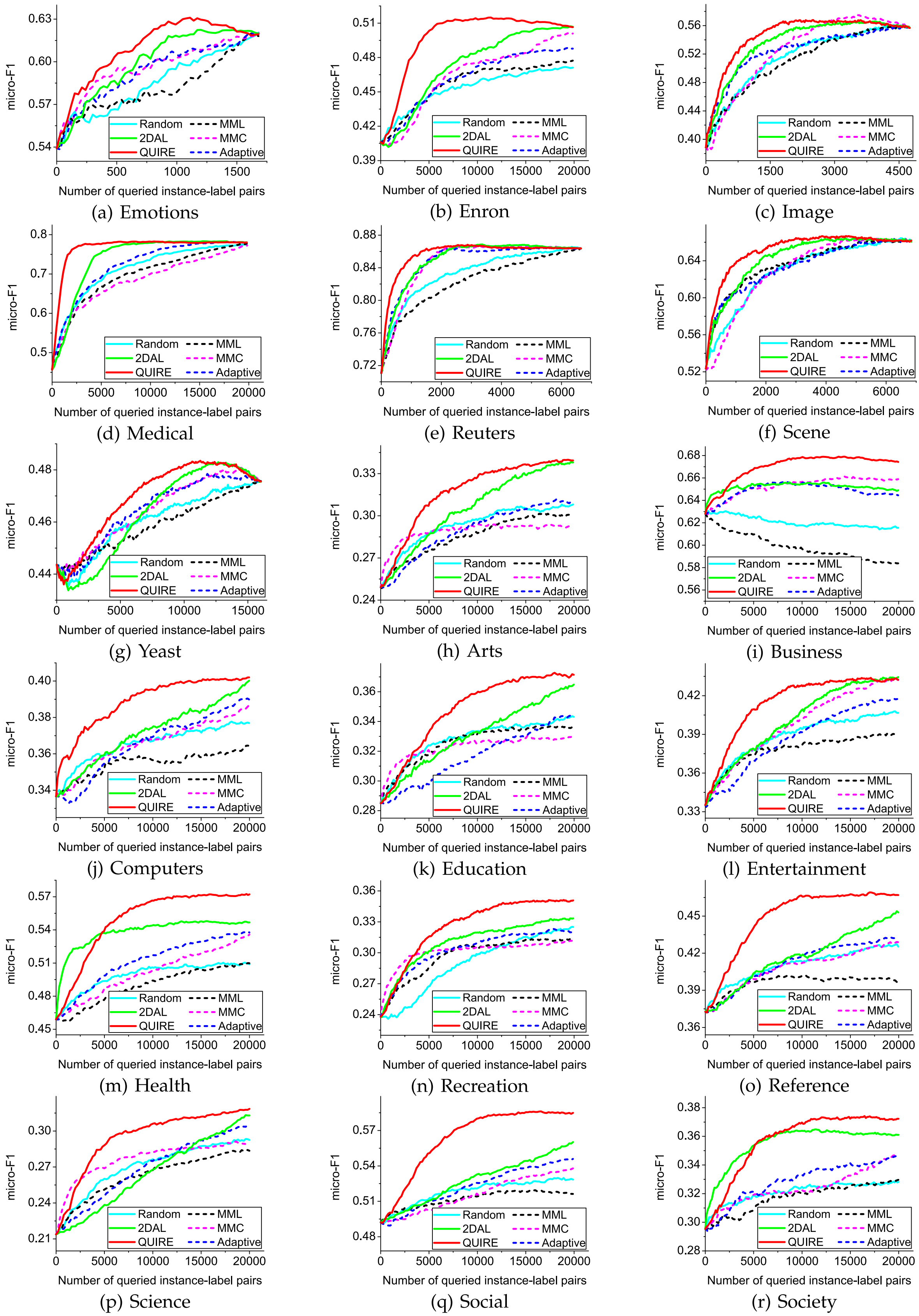


Fig. 3. Comparison on *Micro-F1*.

and recall with equal weights, and can be averaged over instances or labels. Given the large difference of the number of positive instances for different labels, it is less appropriate to equally average over labels. We thus follow [49] to use micro-F1, which first computes the F1-score for each test example and then takes

Authorized licensed use limited to: Huazhong University of Science and Technology. Downloaded on November 02,2020 at 09:30:26 UTC from IEEE Xplore. Restrictions apply.

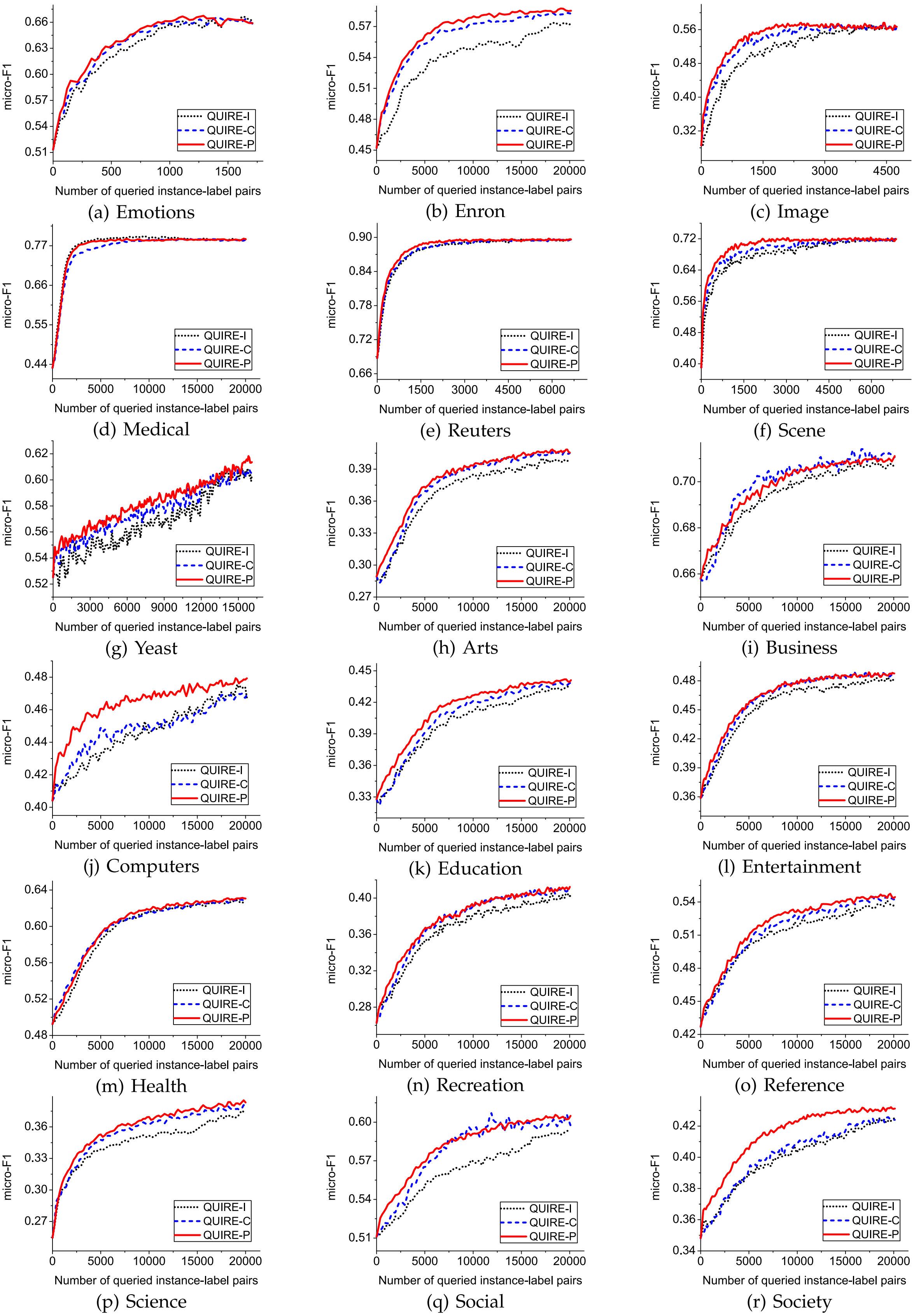


Fig. 4. Micro-F1 curve of QUIRE with different label correlation matrices.

average over all the test examples. It is commonly used in multi-label learning research [23], [49]. A larger micro-F1 indicates a better performance.

5.2.2 Comparison with State-of-the-Art Methods

Since label correlation matrix is usually not easy to obtain, we first study the performance of QUIRE by setting R to the

Authorized licensed use limited to: Huazhong University of Science and Technology. Downloaded on November 02,2020 at 09:30:26 UTC from IEEE Xplore. Restrictions apply.

identity matrix. To be fair, one-versus-rest linear SVM (implemented with LIBLINEAR [14]) is employed as the classification model for evaluating all the compared approaches. For the MMC approach, the regression model is also implemented with LIBLINEAR. For QUIRE, the parameter λ is selected via five-folds cross validation on the initial labeled data from the candidate values $\{1, 10, 100\}$. For the other approaches, parameters are determined in the same way if no values suggested in their literatures.

Fig. 3 shows the performance on micro-F1 with the increasing number of instance-label pair queries. Compared to the baselines, our approach QUIRE achieves the best performance in most cases. In general, we observe that the three methods that use instance-label pair queries (plotted in solid line) are more effective than those that query the entire label vectors for the selected instances (plotted in dashed line). We also observe that for several data sets, the random approach can be more effective than the active learning approaches that solicit all the label assignments for the selected instances. This observation is consistent with the results in [32], suggesting that querying only one chosen label for each selected instance is a more effective strategy.

5.2.3 Study on the Impact of Label Correlation

The previous experiments show that even without exploiting label correlations, QUIRE can outperform state-of-the-art approaches for multi-label active learning. In this section, we study if the performance of QUIRE can be further improved by incorporating the correlation matrix R. Specifically, we employ two simple methods for computing R, i.e., co-occurrence and ϕ-coefficient [43], which are commonly used in the studies of multi-label learning [21], [43]. Since one-versus-rest SVM does not exploit label correlations, it may not be able to clearly show the impact of different label correlations. We thus employ the ensemble of classifier chains (ECC) [33] to train the classification model after each query. ECC is a state-of-the-art multi-label algorithm, which exploits the correlations by linking different labels with a chain of classifiers.

Fig. 4 shows the micro-F1 of three methods with increasing number of instance-label pair queries: (1) QUIRE-I, where R is set to an identity matrix, (2) QUIRE-C, where R is computed based on the co-occurrence between labels, and (3) QUIRE-P, where R is computed based on the ϕ-coefficient. As shown in the figure, QUIRE-C and QUIRE-P usually outperforms QUIRE-I. The advantages of QUIRE-C and QUIRE-P are particularly obvious when the number of labels is large except for data set *medical*, where QUIRE-I is slightly better than the other two methods, possibly due to the relatively poor estimation of label correlation. When comparing the two different estimation of label correlations, ϕ-coefficient tends to be more effective than co-occurrence. We finally note that both ϕ-coefficient and co-occurrence measure the label correlations using only the statistics collected from the training data. We thus expect that the performance of QUIRE can be further improved when the correlation matrix can be estimated more accurately by exploring side information such as domain knowledge.

6 Conclusion

This paper proposes a new active learning approach, QUIRE, for both single-label and multi-label learning, which extends our preliminary research [20]. QUIRE is designed to find unlabeled data that are both informative and representative. It is based on the min-max view of active learning, which provides a systematic way for measuring and combining the informativeness and the representativeness. In the future, we plan to develop a mechanism which allows dynamic and adaptive tradeoff between informativeness and representativeness. In addition, we plan to design multi-label active learning methods that can incorporate the prior knowledge on label correlations.

Acknowledgments

The authors want to thank the anonymous reviewers for their helpful comments and suggestions. This research was supported by the National Fundamental Research Program of China (2014CB340501), the National Science Foundation of China (61333014, 61321491), NSF (IIS-1251031) and ONR Award (N000141210431). Z.-H. Zhou is the corresponding author of this paper.

References

[1] A. Asuncion and D.J. Newman, UCI Machine Learning Repository, 2007.
[2] M.F. Balcan, A.Z. Broder, and T. Zhang, “Margin Based Active Learning,” *Proc. 20th Ann. Conf. Learning Theory*, pp. 35-50, 2007.
[3] M. Belkin, P. Niyogi, and V. Sindhwani, “Manifold Regularization: A Geometric Framework for Learning from Labeled and Unlabeled Examples,” *J. Machine Learning Research*, vol. 7, pp. 2399-2434, 2006.
[4] M. Boutell, J. Luo, X. Shen, and C. Brown, “Learning Multi-Label Scene Classification,” *Pattern Recognition*, vol. 37, no. 9, pp. 1757-1771, 2004.
[5] K. Brinker, “On Active Learning in Multi-Label Classification,” *From Data and Information Analysis to Knowledge Eng.*, pp. 206-213, Springer, 2006.
[6] C.C. Chang and C.J. Lin, “LIBSVM: A Library for Support Vector Machines,” *ACM Trans. Intelligent Systems and Technology*, vol. 2, pp. 27:1-27:27, 2001.
[7] O. Chapelle, B. Schölkopf, and A. Zien, *Semi-Supervised Learning*. MIT Press, 2006.
[8] R. Chattopadhyay, Z. Wang, W. Fan, I. Davidson, S. Panchanathan, and J. Ye, “Batch Mode Active Sampling Based on Marginal Probability Distribution Matching,” *Proc. 18th ACM SIGKDD Int’l Conf. Knowledge Discovery and Data Mining*, pp. 741-749, 2012.
[9] I. Dagan and S.P. Engelson, “Committee-Based Sampling for Training Probabilistic Classifiers,” *Proc. 12th Int’l Conf. Machine Learning*, pp. 150-157, 1995.
[10] S. Dasgupta and D. Hsu, “Hierarchical Sampling for Active Learning,” *Proc. 25th Int’l Conf. Machine Learning*, pp. 208-215, 2008.
[11] P. Donmez, J.G. Carbonell, and P.N. Bennett, “Dual Strategy Active Learning,” *Proc. 18th European Conf. Machine Learning*, pp. 116-127, 2007.
[12] A. Elisseeff and J. Weston, “A Kernel Method for Multi-Labelled Classification,” *Proc. Advances in Neural Information Processing Systems 14*, pp. 681-687, 2001.
[13] A. Esuli and F. Sebastiani, “Active Learning Strategies for Multi-Label Text Classification,” *Proc. 31th European Conf. IR Research on Advances in Information Retrieval*, pp. 102-113, 2009.
[14] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, “Liblinear: A Library for Large Linear Classification,” *The J. Machine Learning Research*, vol. 9, pp. 1871-1874, 2008.
[15] P. Flaherty, M.I. Jordan, and A.P. Arkin, “Robust Design of Biological Experiments,” *Proc. Advances in Neural Information Processing Systems 18*, pp. 363-370, 2005.

Authorized licensed use limited to: Huazhong University of Science and Technology. Downloaded on November 02,2020 at 09:30:26 UTC from IEEE Xplore. Restrictions apply.

[16] Y. Freund, H.S. Seung, E. Shamir, and N. Tishby, "Selective Sampling Using the Query by Committee Algorithm," *Machine Learning*, vol. 28, no. 2/3, pp. 133-168, 1997.

[17] Y. Guo, "Active Instance Sampling via Matrix Partition," *Proc. Advances in Neural Information Processing Systems 23*, pp. 802-810, 2010.

[18] Y. Guo and D. Schuurmans, "Discriminative Batch Mode Active Learning," *Proc. Advances in Neural Information Processing Systems 20*, pp. 593-600, MIT Press, 2007.

[19] S.C.H. Hoi, R. Jin, J. Zhu, and M.R. Lyu, "Semi-Supervised SVM Batch Mode Active Learning for Image Retrieval," *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 1-7, 2008.

[20] S.-J. Huang, R. Jin, and Z.-H. Zhou, "Active Learning by Querying Informative and Representative Examples," *Proc. Advances in Neural Information Processing Systems 23*, pp. 892-900, 2010.

[21] S.-J. Huang, Y. Yu, and Z.-H. Zhou, "Multi-Label Hypothesis Reuse," *Proc. 18th ACM SIGKDD Conf. Knowledge Discovery and Data Mining*, pp. 525-533, 2012.

[22] S.-J. Huang and Z.-H. Zhou, "Active Query Driven by Uncertainty and Diversity for Incremental Multi-Label Learning," *Proc. 13th IEEE Int'l Conf. Data Mining*, pp. 1079-1084, 2013.

[23] C.-W. Hung and H.-T. Lin, "Multi-Label Active Learning with Auxiliary Learner," *Proc. Third Asian Conf. Machine Learning*, pp. 315-330, 2011.

[24] B. Klimt and Y. Yang, "Introducing the Enron Corpus," *Proc. First Conf. Email and Anti-Spam*, 2004.

[25] D.D. Lewis and J. Catlett, "Heterogeneous Uncertainty Sampling for Supervised Learning," *Proc. 11th Int'l Conf. Machine Learning*, pp. 148-156, 1994.

[26] D.D. Lewis and W.A. Gale, "A Sequential Algorithm for Training Text Classifiers," *Proc. 17th Ann. Int'l ACM-SIGIR Conf. Research and Development in Information Retrieval*, pp. 3-12, 1994.

[27] C.-L. Li, C.-S. Ferng, and H.-T. Lin, "Active Learning with Hinted Support Vector Machine," *J. Machine Learning Research-Proceedings Track*, vol. 25, pp. 221-235, 2012.

[28] X. Li and Y. Guo, "Active Learning with Multi-Label SVM Classification," *Proc. 23rd Int'l Joint Conf. Artificial Intelligence*, 2013.

[29] X. Li, L. Wang, and E. Sung, "Multilabel SVM Active Learning for Image Classification," *Proc. Int'l Conf. Image Processing*, vol. 4, pp. 2207-2210, 2004.

[30] H.T. Nguyen and A.W.M. Smeulders, "Active Learning Using Pre-Clustering," *Proc. 21st Int'l Conf. Machine Learning*, pp. 623-630, 2004.

[31] J. Petterson and T. Caetano, "Submodular Multi-Label Learning," *Proc. Advances in Neural Information Processing Systems 24*, pp. 1512-1520, 2011.

[32] G.-J. Qi, X.-S. Hua, Y. Rui, J. Tang, and H.-J. Zhang, "Two-Dimensional Active Learning for Image Classification," *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 1-8, 2008.

[33] J. Read, B. Pfahringer, G. Holmes, and E. Frank, "Classifier Chains for Multi-Label Classification," *Machine Learning*, vol. 85, no. 3, pp. 333-359, 2011.

[34] R. Rifkin, G. Yeo, and T. Poggio, "Regularized Least Squares Classification," *Advances in Learning Theory: Methods, Model and Applications, NATO Science Series III: Computer and Systems Sciences*, pp. 131-154, IOS Press, 2003.

[35] N. Roy and A. Mccallum, "Toward Optimal Active Learning through Sampling Estimation of Error Reduction," *Proc. 18th Int'l Conf. Machine Learning*, pp. 441-448, 2001.

[36] F. Sebastiani, "Machine Learning in Automated Text Categorization," *ACM Computing Surveys*, vol. 34, no. 1, pp. 1-47, 2002.

[37] B. Settles, "Active Learning Literature Survey," Computer Sciences Technical Report 1648, Univ. of Wisconsin–Madison, 2009.

[38] H.S. Seung, M. Opper, and H. Sompolinsky, "Query by Committee," *Proc. Int'l Workshop Computational Learning Theory*, pp. 287-294, 1992.

[39] M. Singh, E. Curran, and P. Cunningham, "Active Learning for Multi-Label Image Annotation," *Proc. 19th Irish Conf. Artificial Intelligence and Cognitive Science*, 2008.

[40] L. Sun, S. Ji, and J. Ye, "Hypergraph Spectral Learning for Multi-Label Classification," *Proc. 14th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining*, pp. 668-676, 2008.

[41] S. Tong and D. Koller, "Support Vector Machine Active Learning with Applications to Text Classification," *Proc. 17th Int'l Conf. Machine Learning*, pp. 999-1006, 2000.

[42] K. Trohidis, G. Tsoumakas, G. Kalliris, and I. Vlahavas, "Multi-Label Classification of Music into Emotions," *Proc. Ninth Int'l Conf. Music Information Retrieval*, pp. 325-330, 2008.

[43] G. Tsoumakas, A. Dimou, E. Spyromitros, and V. Mezaris, "Correlation-Based Pruning of Stacked Binary Relevance Models for Multi-Label Learning," *Proc. First Int'l Workshop Learning from Multi-Label Data*, pp. 101-116, 2009.

[44] G. Tsoumakas, I. Katakis, and I. Vlahavas, "Mining Multi-Label Data," *Data Mining and Knowledge Discovery Handbook*, pp. 667-685, Springer, 2010.

[45] N. Ueda and K. Saito, "Parametric Mixture Models for Multi-Labeled Text," *Proc. Advances in Neural Information Processing Systems 15*, pp. 721-728, 2002.

[46] Z. Wang and J. Ye, "Querying Discriminative and Representative Samples for Batch Mode Active Learning," *Proc. 19th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining*, pp. 158-166, 2013.

[47] Z. Xu, K. Yu, V. Tresp, X. Xu, and J. Wang, "Representative Sampling for Text Classification Using Support Vector Machines," *Proc. 25th European Conf. Information Retrieval Research*, pp. 393-407, 2003.

[48] R. Yan, J. Tešić, and J.R. Smith, "Model-Shared Subspace Boosting for Multi-Label Classification," *Proc. 13th ACM SIGKDD Conf. Knowledge Discovery and Data Mining*, pp. 834-843, 2007.

[49] B. Yang, J.-T. Sun, T. Wang, and Z. Chen, "Effective Multi-Label Active Learning for Text Classification," *Proc. 15th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining*, pp. 917-926, 2009.

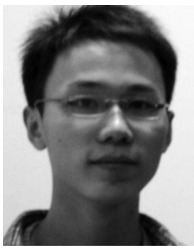
[50] K. Yu, J. Bi, and V. Tresp, "Active Learning via Transductive Experimental Design," *Proc. 23rd Int'l Conf. Machine Learning*, pp. 1081-1088, 2006.

[51] M.-L. Zhang and K. Zhang, "Multi-Label Learning by Exploiting Label Dependency," *Proc. 16th ACM SIGKDD Conf. Knowledge Discovery and Data Mining*, pp. 999-1007, 2010.

[52] M.-L. Zhang and Z.-H. Zhou, "ML-kNN: A Lazy Learning Approach to Multi-Label Learning," *Pattern Recognition*, vol. 40, no. 7, pp. 2038-2048, 2007.

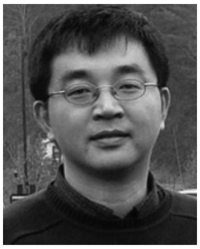
[53] M.-L. Zhang and Z.-H. Zhou, "A Review on Multi-Label Learning Algorithms," *IEEE Trans. Knowledge and Data Eng.*, doi. ieeecomputersociety.org/10.1109/TKDE.2013.39.

[54] Z.-H. Zhou, M.-L. Zhang, S.-J. Huang, and Y.-F. Li, "Multi-Instance Multi-Label Learning," *Artificial Intelligence*, vol. 176, no. 1, pp. 2291-2320, 2012.



Sheng-Jun Huang received the BSc degree from Nanjing University, China, in 2008. He is currently working toward the PhD degree in the Department of Computer Science & Technology of Nanjing University. His main research interests include machine learning and data mining. He received the Microsoft Fellowship Award in 2011 and the Best Poster award at the 18th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD) in 2012.

Authorized licensed use limited to: Huazhong University of Science and Technology. Downloaded on November 02,2020 at 09:30:26 UTC from IEEE Xplore. Restrictions apply.



Rong Jin received the PhD degree from Carnegie Mellon University in 2003. He is a professor in the Computer and Science Engineering Department at Michigan State University. He has been working in the areas of statistical machine learning and its application to information retrieval. He has extensive research experience in a variety of machine learning algorithms such as conditional exponential models, support vector machine, boosting and optimization for different applications including information retrieval. He is an associate editor of *ACM Transactions on Knowledge Discovery from Data*, and received the National Science Foundation (NSF) Career Award in 2006. He received the Best Paper award from Conference of Learning Theory (COLT) in 2012. He is a member of the IEEE.

Zhi-Hua Zhou (S'00-M'01-SM'06-F'13) received the BSc, MSc, and PhD degrees in computer science from Nanjing University, China, in 1996, 1998, and 2000, respectively, all with the highest honors. He joined the Department of Computer Science & Technology at Nanjing University as an assistant professor in 2001, and is currently a professor and the Director of the LAMDA group. His research interests are mainly in artificial intelligence, machine learning, data mining, pattern recognition, and multimedia information retrieval. In these areas he has published more than 100 papers in leading international journals or conference proceedings, and holds 12 patents. He has won various awards/honors including the IEEE CIS Outstanding Early Career Award, the National Science & Technology Award for Young Scholars of China, the Fok Ying Tung Young Professorship Award, the Microsoft Young Professorship Award and nine international journals/conferences paper or competition awards. He is an executive editor-in-chief of the *Frontiers of Computer Science*, associate editor-in-chief of the *Chinese Science Bulletin*, an associate editor or editorial boards member of the *ACM Transactions on Intelligent Systems and Technology*, *IEEE Transactions on Neural Networks and Learning Systems*, etc. He served as an associate editor for *IEEE Transactions on Knowledge and Data Engineering* (2008-2012) and *Knowledge and Information Systems* (2003-2008). He is the founder and the Steering Committee chair of ACML, and Steering Committee member of PAKDD and PRICAI. He serves/ed as the general chair/co-chair of ACML12, ADMA12, PCM13, PAKDD14, program chair/co-chair of PAKDD07, PRICAI08, ACML09, SDM13, etc., Workshop chair/co-chair of KDD12 and ICDM14, tutorial chair/co-chair of KDD13 and CIKM14, and the program vice chair or the area chair of various conferences such as ICML, IJCAI, AAAI, ICPR, etc. He is the chair of the Machine Learning Technical Committee of the Chinese Association of Artificial Intelligence, the chair of the Artificial Intelligence & Pattern Recognition Technical Committee of the China Computer Federation, the vice chair of the Data Mining Technical Committee of IEEE Computational Intelligence Society and the chair of the IEEE Computer Society Nanjing Chapter. He is a fellow of the IEEE, IAPR fellow, IET/IEE fellow, and an ACM distinguished scientist.

▷ **For more information on this or any other computing topic, please visit our Digital Library at** www.computer.org/publications/dlib.

Authorized licensed use limited to: Huazhong University of Science and Technology. Downloaded on November 02,2020 at 09:30:26 UTC from IEEE Xplore. Restrictions apply.