Multi-label Active Learning Based on Maximum Correntropy Criterion: Towards Robust and Discriminative Labeling

Zengmao Wang[1], Bo Du[1(✉)], Lefei Zhang[1], Liangpei Zhang[2], Meng Fang[3], and Dacheng Tao[4]

[1] State Key Laboratory of Software Engineering, School of Computer, Wuhan University, Wuhan, China
{kingmao,remoteking,zhanglefei}@whu.edu.cn
[2] State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China
zlp62@whu.edu.cn
[3] Department of Computing and Information Systems, University of Melbourne, Parkville, Australia
meng.fang@unimelb.edu.au
[4] QCIS and FEIT, University of Technology Sydney, Sydney, NSW 2007, Australia
dacheng.tao@uts.edu.au

Abstract. Multi-label learning is a challenging problem in computer vision field. In this paper, we propose a novel active learning approach to reduce the annotation costs greatly for multi-label classification. State-of-the-art active learning methods either annotate all the relevant samples without diagnosing discriminative information in the labels or annotate only limited discriminative samples manually, that has weak immunity for the outlier labels. To overcome these problems, we propose a multi-label active learning method based on Maximum Correntropy Criterion (MCC) by merging uncertainty and representativeness. We use the the labels of labeled data and the prediction labels of unknown data to enhance the uncertainty and representativeness measurement by merging strategy, and use the MCC to alleviate the influence of outlier labels for discriminative labeling. Experiments on several challenging benchmark multi-label datasets show the superior performance of our proposed method to the state-of-the-art methods.

Keywords: Multi-label learning · Active learning · Correntropy · Robust

1 Introduction

Active learning has been widely used in computer visions to address the samples imbalance problem that the available labeled data is much less than the unlabeled data [18,35]. It is an iterative loop to find the most valuable samples for the oracle to label, and gradually improves the model generalization ability until the convergence condition is satisfied [39]. There are two motivations behind the design of

B. Leibe et al. (Eds.): ECCV 2016, Part III, LNCS 9907, pp. 453–468, 2016.
DOI: 10.1007/978-3-319-46487-9_28

a practical active learning algorithm, namely, uncertainty and representativeness [8,15]. Uncertainty is to improve the models' generalization ability and representativeness is to prevent the bias of the models.

Among all the active learning based tasks, multi-label classification, which aims to assign each object with multiple labels, may be the most difficult and costly one [10,17,42]. In current research, active learning for multi-label learning has become even more important, reducing the costs of the various multi-label tasks [6,7,38,41]. State-of-the-art multi-label active learning can be classified into three categories based on the query function used to select the valuable samples. The first category relies on the labeled data to design a query function with uncertainty [25,28]. In such methods, the design of the query function ignores the latent structural information in the large-scale unlabeled data, leading to a serious sample bias and an undesirable performance. To eliminate this problem, the second category, which depends on the representativeness, has been developed [26]. In these approaches, the structural information of the unlabeled data is elaborately considered, but the discriminative (uncertain) information is discarded. Therefore, a large number of samples would be required before an optimal boundary is found. Since utilizing either the uncertainty criterion or the representativeness criterion may not achieve a desirable performance, the third category which combines both criteria borns naturally and it can effectively solve these problems [8,15]. However, the approaches in the third category are either heuristic in designing the specific query criterion or ad hoc in measuring the uncertainty and representativeness of the samples. The uncertainty still just relies on the limited labeled data. Most importantly, previous works ignore the outlier labels that exist in multi-label classification when designing a query model for active learning.

However, the outlier labels have significant influence on the measurement of uncertainty and representativeness in multi-label learning. In the following, we will discuss the outlier label and its negative influence on the measurement of uncertainty and representativeness in details.

Figure 1 shows a simple example about the influence of outlier labels. As the input, we annotate the image with three labels, namely tree, elephant and lion. Hence, the feature of image is combined with three parts, the feature of tree,

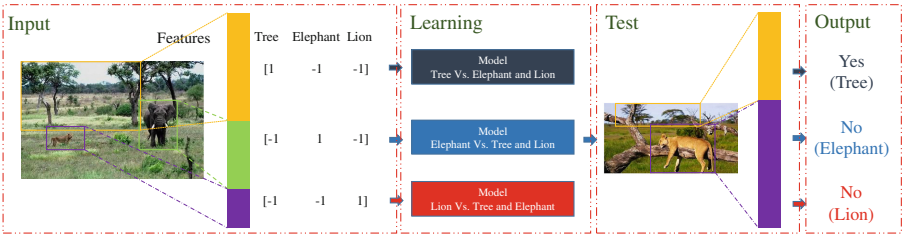


Fig. 1. The influence of outlier label in the learning process.

the feature of elephant and the feature of lion. Intuitively, in the image feature, the feature of tree is much more than elephant and lion, and the feature of lion is the least. If we use the image with the three labels to learn a lion/ non-lion binary classification model, the model would actually depend on the trees and elephants features rather than the lions. Thus it would be a biased model for classifying the lion and the non-lions. Given the test image where a lion covers the most regions in the image, the trained model would not recognize it. If we use such a model to measure the uncertainty in active learning, it may cause error measurement for images with lion label. We name the lion label in the input image as an outlier label.

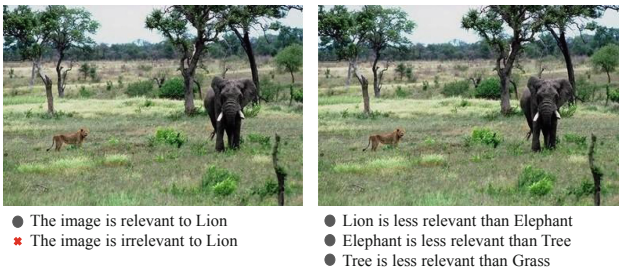


Fig. 2. The interface of two properties for outlier labels. Left: The outlier label (Lion) is relevant to the image; right: the outlier (Lion) is much less relevant to the image than the most relevant label (Tree) is.

Furthermore, we present the formal definition of the outlier label. Denote x, y_1 and y_2 as the selected instance and two relevant labels, respectively. Define y_1 as the outlier label, if it has two properties. The first one is that y_1 is a relevant label to the instance x, and the second is that y_1 is much less relevant to x than y_2 is. If the trained model could determine whether y_1 or y_2 is the outlier label to x, it would be very useful to build a promising model for a better query. Figure 2 shows the two properties. The definition of the outlier label is consistent with the fact that, given an image, some labels relevance to it is apparent, which can be recognized at first glance by the oracle, and some labels relevance is veiled, which may need much effort for the oracle to label. The definition of outlier label is also consistent with the query types proposed in [14]. For two multi-label images, if they have the same labels, but the outlier label is different in their labels, this may lead to the features of the two image have a large difference, therefore, it is very hard to diagnose the similarity between two instances with outlier labels. In Fig. 3, we provide a simple example to show such a problem, and we present the similarity between the sift features with Gaussian kernel [1,22,30,32], and the labels similarity based on MCC. Intuitively, the similarity between image 1 and image 2 should be larger than the similarity between image 2 and image 3, since the labels in image 1 and image 2 are exactly the same. However, the result is opposite when the similarity is measured with their sift features. This because

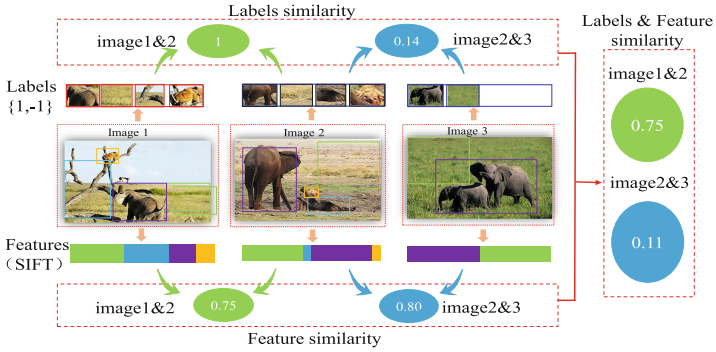


Fig. 3. The influence of the outlier labels for the measurement of similarity

the outlier label is lion in image 1, tree trunk and lion are two outlier labels in image 2, and the different outlier labels largely increase the difference between the features of the two images. In summary, the measurement of uncertainty and representativeness would be deteriorated with the outlier labels.

To address the above problems, in this paper, we propose the robust multi-label active learning (RMLAL) algorithm, which effectively combines the uncertainty and representativeness based on the MCC [40].

As to robustness, the correntropy has proved promising in information theoretic learning (ITL) and can efficiently handle the large outliers [40]. In traditional active learning algorithms, the mean square error (MSE) cannot easily control the large errors caused by the outliers [12,19,23,27,29,36]. We therefore replace the MSE criterion with the MCC in the proposed formulation with a minimum margin model. In this way, the proposed method is able to eliminate the outlier samples, making the query function more robust.

As to discriminative labeling, we use the MCC to measure the loss between the true label and the prediction label. MCC can improve the most discriminative information and suppress the little useless information or unexpected information. Hence, with MCC in the proposed method, if the label is not an outlier label, it will play an important role in the query model construction. Otherwise, the model will decrease the influence of the outlier label to measure the uncertainty. Then the discriminative labels effects are improved and the outlier labels are suppressed, and the discriminative labeling can be achieved.

For representativeness, we mix the prediction labels of unlabeled data with the MCC as the representativeness. As is shown in Fig. 3, although the samples have the same labels, their outlier labels are different, making their features distinguishing. If we just use the corresponding features to measure the similarity, it will lead to a wrong diagnosis. Hence, we propose to use the combination of labels and sample similarity to define the consistency between the labels and samples. With different space measurement making up for each other [33,34,37], the combination makes the measurement of representativeness more general. To decrease

the computational complexity of the proposed method, the half-quadratic optimization technique is adopted to optimize the MCC. The contributions of our work can be summarized as follows:

- To the best of our knowledge, it is the first work to focus on the outlier labels in multi-label active learning. We find a robust and effective query model for multi-label active learning.
- The prediction labels of unlabeled data and the labels of labeled data are utilized with MCC to merge the uncertain and representative information, deriving an approach to make the uncertain information more precise.
- The proposed representative measurement considers labels similarity by MCC. It can effectively handle the outlier labels and makes the similarity more accuracy for multi-label data, and also provides a way to merge representativeness into uncertainty.

The rest of the paper is organized as following: Sect. 2 briefly introduces the related works. Then Sect. 3 defines and discusses a new objective for robust multi-label active learning and proposes an algorithm based on half-quadratic optimization. Section 4 evaluates our method on several benchmark multi-label data sets. Finally, we summarize the paper in Sect. 5.

2 Related Works

Since multi-label problem is universal in the real world, it has drawn great interests in many fields. For a multi-label object, it needs an oracle to consider all the relevant labels, leading to the labeling of multi-label tasks is more costly than single label learning, however, the research of active learning on multi-label is still less.

In multi-label learning, one instance is corresponding to more than one labels. To solve a multi-label problem, it is a direct way to convert it to several binary problems [21,31]. In these approaches, the uncertainty is measured for each label, and then a combining strategy is adopted to measure the uncertainty of one instance. [21] trained a probabilistic binary logistic regression classifier with different levels, and combined them with level switching strategy for adaptive selection. [31] converted the SVM margin to a probability score to select the instance for query. Recently, [26] selected the valuable instances by minimizing the Expected Error Reduction. Other works have done by combining the informativeness and representativeness together for a better query [8,20]. [20] combined the label cardinality inconsistency and the separation margin with a tradeoff parameter. [8] incorporated the data distribution in the selection process by using the appropriate dissimilarity between pairs of samples with sparse modeling representative selection for query. All the above algorithms were designed to query all the labels of the query instances. Another approaches have been developed to query the label-instance pairs with relevant label and instance at each iteration [14,16]. [14] queried the instance with relevant labels based on the types. [16] selected label-instance pairs based on a label ranking model. In these

approaches, some important labels may be lost. In this study, considering the combination of informativeness and representativeness is very effective in active learning, we adopt this strategy.

No matter selecting the instance by all the labels or by the label-instance pairs, most of the active learning algorithms only selected the uncertain instance based on very limited samples, and ignored the labels information. For example, given all the labels to one instance, if the outlier labels are too much in label ranking, such instance may decrease the performance of the task. Moreover, given the relevant labels to one instance, some relevant labels may be lose with the limited query labels. To address these problems, we use the prediction labels of unlabeled data to enhance the uncertain measurement and adopt the MCC to consider the much relevant labels as much as possible except the outlier labels. As far to our knowledge, it is the first time to adopt the MCC in multi-label active learning with data labels for query.

3 Methodology

Suppose we are given a multi-label data set $D = \{x_i, x_2, \ldots, x_n\}$ with n samples and C possible labels for each sample. Initially, we label l samples in D. Without loss of generality, we denote the l labeled samples as set $L = \{(x_1, y_1), (x_2, y_2), \ldots, (x_l, y_l)\}$, where $y_i = (y_{i1}, y_{i2}, \ldots, y_{iC})$ is the labels set for sample x_i, with $y_{ik} \in \{-1, 1\}$; and the remaining $u = n - l$ unlabeled samples are denoted as set $U = \{x_{l+1}, x_{l+2}, \ldots, x_{l+u}\}$. It is the candidate set for active learning. Moreover, we denote x_q as the query sample in the active learning process. In each iteration, we select $x_q \in U$. And we use the bold symbol to denote the matrix or vector. In the following discussion, the symbols are used as above.

3.1 Maximum Correntropy Criterion

In multi-label classification tasks, the outlier labels pose a great challenge to train a precise classifier, mainly due to the unpredictable nature of the errors (bias) caused by these outliers. In active learning, in particular, the limited labeled samples with outliers easily lead to great bias. Since in active learning the supervised information is limited, it is hard to avoid the influence of the outlier labels when building the supervised model. This directly leads to the bias of uncertain information, furthermore makes the query instances are undesirable or even leads to bad performance.

Recently, the concept of correntropy was firstly proposed in ITL and it had drawn much attention in the signal processing and machine learning community for robust analysis, which can effectively handle the outliers [13]. In fact, correntropy is a similarity measure between two arbitrary random variables a and b [13], defined by

$$\hat{V}_\sigma(a, b) = E[K_\sigma(a, b)] \tag{1}$$

where $K_\sigma(\cdot)$ is the kernel function and $E[\cdot]$ is the expectation operator. We can observe that the definition of correntropy bases the kernel method, so it also has the same advantages that the kernel technique owns. However, different from the conventional kernel based methods, correntropy works independently with pairwise samples and has a strong theoretical foundation. With such a definition, the properties of correntropy are symmetric, positive and bounded.

Since the joint probability density function of a and b in practice is unknown, and the available data $\{a_i, b_i\}_{i=1}^n$ are usually finite, the sample estimator of correntropy is usually adopted by

$$\hat{V}_\sigma(a,b) = E[K_\sigma(a,b)] \tag{2}$$

where $K_\sigma(x_1, x_2) = exp(-||x_1 - x_2||^2)/2\sigma^2)$. According to [13], the correntropy between a and b is given by

$$\max_{p'} \frac{1}{n} \sum_{i=1}^{n} K_\sigma(a_i, b_i) \tag{3}$$

The objective function (3) is called maximum correntropy criterion (MCC) [13], where p' is the auxiliary parameter to be specified in Proposition 1. Compared with mean square error (MSE), which is a global metric, the correntropy is a local metric. That means the correntropy value is mainly determined by the kernel function along the line A = B [11].

3.2 The Proposed Approach

Usually, the uncertainty is measured according to the labeled data whereas the representativeness according to the unlabeled data. In this paper, we propose a novel approach to merge the uncertainty and representativeness of instances in active learning. Minimum margin is the most popular and direct approach to measure the uncertainty, which chooses the unlabeled sample by its prediction uncertainty [15]. Let f^* be the classifier that is trained by the labeled samples, and the sample x_q that we want to query in the unlabeled data based on the margin can be found as follows:

$$x_q = \arg\min_{x_i \in U} |f^*(x_i)| \tag{4}$$

Generally, with the labeled samples, we can find a classification model f^* for a binary class problem in supervised approach with the following loss function:

$$f^* = \arg\min_{f \in \mathcal{H}} \sum\nolimits_{x_i \in L} \ell(Y_i, f(x_i)) + \lambda \|f\|_{\mathcal{H}}^2 \tag{5}$$

where $\mathcal{H}$ is a reproducing kernel Hilbert space endowed with kernel function $K(\cdot)$, $\ell(\cdot)$ is the loss function and Y_i belongs to $\{1, -1\}$. Following the works of [15], the criterion of the minimum margin can be written as

$$x_q = \arg\min_{x_j \in U} \max_{Y_j \pm 1} \min_{f \in \mathcal{H}} \sum_{x_i \in L} \ell(Y_i, f(x_i)) + \lambda \|f\|_{\mathcal{H}}^2 + \ell(Y_j, f(x_j)) \tag{6}$$

Y_j is a pseudo label for the unlabeled sample x_j. Since it is a binary class problem, Y_j is 1 or -1. Hence, we define $Y_j = -sign(f(x_j))$. In previous works, the loss function is adopted with quadratic loss for MSE, but it is not robust for the occasion of outliers. To overcome this problem, considering the properties of MCC, we introduce the MCC as the loss function. Different from MSE by minimizing the loss to solve minimization problem, MCC solves the minimization problem by maximizing the loss, presented by

$$\underset{x_q \in U, Y_q = \pm 1, f \in \mathcal{H}}{\arg\max} \sum_{x_i \in L} \exp\left(-\frac{\|Y_i - f(x_i)\|^2}{2\sigma^2}\right) - \lambda \|f\|_{\mathcal{H}}^2 + \exp\left(-\frac{\|Y_q - f(x_q)\|^2}{2\sigma^2}\right) \tag{7}$$

where σ is the kernel width. Following the minimum margin approach, the objective function (7) is equal to (4). In our work, we extend multi-label classification as several binary classification problems with label correlation [15]. For simple, we assume the label correlation is independent by learning one classifier for each label independently. Then, we use the summation as minimum margin in multi-label learning and use f_i as the classifier between i^{th} label and the other labels. The multi-label active learning to query the sample with minimum margin approach based on MCC with the worst case is given by

$$\begin{aligned} \mathcal{L}(x_q, f, L) = & \underset{x_j \in U, f_k \in \mathcal{H}: k = \{1,2,..C\}}{\arg\max} \sum_{x_i \in L} \sum_{k=1}^{C} \exp\left(-\frac{\|y_{ik} - f_k(x_i)\|^2}{2\sigma^2}\right) \\ & -\lambda \sum_{k=1}^{C} \|f_k\|_{\mathcal{H}}^2 + \sum_{k=1}^{C} \exp\left(-\frac{\left(1 + 2|f_k(x_q)| + f_k(x_q)^2\right)}{2\sigma^2}\right) \end{aligned} \tag{8}$$

The labeled samples in L are very limited, so that it is very important to utilize the unlabeled data to enhance the performance of active learning. Since the labels of the unlabeled data are unknown, it is hard to add the unlabeled data in the supervised model. For the purpose to enhance the uncertain information, we merge the representative information into the uncertain information by prediction labels of unlabeled data. However, the current similarity is difficult to use the unlabeled data to enhance the uncertain information just with features. To overcome this problem, and considering the outlier labels influence, we take the prediction labels of unlabeled data into consideration for similarity measurement. We define a novel consistency between labels and sample similarity with sample-label pairs based on MCC as

$$s\left((x_i, y_i), (x_j, y_j)\right) = \exp\left(-\frac{\|y_i - y_j\|_2^2}{2\sigma^2}\right) w_{ij} \tag{9}$$

where w_{ij} is the similarity between two samples with kernel function. Let $\boldsymbol{S} = [s_{ij}]^{u \times u}$ denote the symmetric similarity matrix for the unlabeled data, and s_{ij} is the consistency between x_i and x_j sample-label pairs points. With such a consistency matrix, the representativeness is to find the sample that can well

represent the unlabeled data set. To do so, [8] proposed a convex optimization framework by introducing variables $p_{ij} \in [0,1]$ which indicates the probability that x_i represents x_j. In our consistency measurement based on MCC, if x_i can represent the point x_j, and it cannot represent the point x_t, there will be $s_{ij} \gg s_{it}$. Such a consistency measurement has already made the difference between representatives and non-representatives large. Therefore, we define that if x_i is the representative one, the probabilities $p_{ij}, j = 1, 2, ..u$ between x_i and the other unlabeled samples are 1, otherwise $p_{ij}, j = 1, 2, ..u$ are 0. Equally, we define $\boldsymbol{d} = [d_{ij}]^{u \times l}$ and $\boldsymbol{z} = [z_{ij}]^{u \times l}$ as the consistency matrix and probability between the unlabeled data and the labeled data respectively. By querying a desirable sample, which can not only represent the unlabeled data and but also not overlap the information in labeled data, we maximize the expectation operator and use a tradeoff parameter β to measure and balance the representative information in unlabeled data and labeled data

$$E\left[x_q, U, L\right] = \max_{x_q} \sum_{x_q \in U} \left[\left(\frac{1}{u} \sum_{x_j \in U} s_{qj} p_{qj} \right) - \beta \left(\frac{1}{l} \sum_{x_j \in L} d_{qj} z_{qj} \right) \right] \tag{10}$$

In current research, it has proved that the combination between uncertainty and representativeness is very effective in active learning [8,15]. In our approach, we also combine them with a tradeoff parameter, given by

$$\mathcal{L}(x_q, f, L) + \beta_0 E[x_q, U, L] \tag{11}$$

To merge the representative part into uncertain part, we use the prediction labels of unlabeled data. For each classifier f_k, we define $f_k(x)$ with a linear regression model in the kernel space as $f_k(x) = \omega_k^T \Phi(x)$ for each label, where $\Phi(x)$ is the feature mapping to the kernel space. In (11), the specific point x_q can be queried from the unlabeled data, but exhaustive search is not feasible due to the exponential nature of the search space. To solve such a problem, we use the numerical optimization-based techniques. An indicator vector $\boldsymbol{\alpha}$ is introduced, which is a binary vector with u length. Each entry α_j denotes whether the corresponding sample x_j is queried as the query sample. If x_j is queried as x_q, α_j is 1, otherwise, α_j is 0. Then the objective function can be defined as

$$\begin{aligned}
\underset{\omega;\alpha^T 1=1,\alpha_i \in \{0,1\}}{\arg\max} & \sum\nolimits_{x_i \in L} \sum_{k=1}^{C} \exp\left(-\frac{\left\| y_{ik} - \omega_k^T \Phi\left(x_i\right) \right\|^2}{2\sigma^2} \right) - \lambda \sum_{k=1}^{C} \left\| \omega_k \right\|^2 \\
& + \sum_{x_j \in U} \alpha_j \sum_{k=1}^{C} \exp\left(-\frac{\left(1 + 2\left| \omega_k^T \Phi\left(x_j\right) \right| + \left(\omega_k^T \Phi\left(x_j\right) \right)^2 \right)}{2\sigma^2} \right) \\
+\beta_1 \sum_{x_j \in U} \alpha_j \left(\frac{1}{u} \right) & \sum_{x_i \in U} \exp\left(-\frac{\left\| \boldsymbol{\omega}^T [\boldsymbol{I} \otimes \Phi\left(x_j\right)] - \boldsymbol{\omega}^T [\boldsymbol{I} \otimes \Phi\left(x_i\right)] \right\|_2^2}{2\sigma^2} \right) w_{ji} \\
& -\beta_2 \sum_{x_j \in U} \alpha_j \left(\frac{1}{l} \right) \sum_{x_i \in L} \exp\left(-\frac{\left\| \boldsymbol{\omega}^T [\boldsymbol{I} \otimes \Phi\left(x_j\right)] - y_i \right\|_2^2}{2\sigma^2} \right) w_{ji}
\end{aligned} \tag{12}$$

where $\boldsymbol{\omega} = \{\omega_1, \omega_2, \ldots, \omega_C\}$ is the multi-label classifier. $\boldsymbol{I}$ is the identify matrix of size $C \times C$, and $\otimes$ is the kronecker product between matrices. Although the objective function (12) is neither convex nor linear, we derive an iterative algorithm based on half-quadratic technique [11,13] with the alternating optimization strategy [2] to solve it efficiently. Based on the theory of convex conjugated functions [3], we can easily derive the following proposition [40].

Proposition 1. *A convex conjugate function φ is exiting to make sure* $g(x) = \exp\left(-\frac{x^2}{2\sigma^2}\right) = \max_{p'}\left(p'\frac{\|x\|^2}{\sigma^2} - \varphi(p')\right)$ *where p' is the auxiliary variable, and with a fixed x, $g(x)$ reaches the maximum value at $p' = -g(x)$.*

According to the Proposition 1, the objective function (12) can be formulated as

$$\begin{aligned}\arg\min_{\omega;\alpha^T 1=1,\alpha_i\in\{0,1\}} & \sum\nolimits_{x_i\in L}\sum_{k=1}^{C}\left[m_{ik}\left\|y_{ik}-\omega_k^T\Phi(x_i)\right\|^2\right]+\lambda\sum_{k=1}^{C}\|\omega_k\|^2 \\ & +\sum_{x_j\in U}\alpha_j\sum_{k=1}^{C}\left[n_{jk}\left(1+2\left|\omega_k^T\Phi(x_j)\right|+\left(\omega_k^T\Phi(x_j)\right)^2\right)\right] \\ -\beta_1\sum_{x_j\in U}\alpha_j\left(\frac{1}{u}\right)\sum_{x_i\in U} & h_{ji}\left\|\boldsymbol{\omega}^T[\boldsymbol{I}\otimes\Phi(x_j)]-\boldsymbol{\omega}^T[\boldsymbol{I}\otimes\Phi(x_i)]\right\|_2^2 w_{ji} \\ & +\beta_2\sum_{x_j\in U}\alpha_j\left(\frac{1}{l}\right)\sum_{x_i\in L} v_{ji}\left\|\boldsymbol{\omega}^T[\boldsymbol{I}\otimes\Phi(x_j)]-y_i\right\|_2^2 w_{ji}\end{aligned} \tag{13}$$

where m_{ik}, n_{jk}, h_{ji}, and v_{ji} are the auxiliary variables, with

$$m_{ik} = \exp\left(-\frac{\left\|y_{ik}-\omega_k^T\Phi(x_i)\right\|^2}{2\sigma^2}\right), x_i \in L, y_{ik} \in y_i$$

$$n_{jk} = \exp\left(-\frac{\left(1+2\left|\omega_k^T\Phi(x_j)\right|+\left(\omega_k^T\Phi(x_j)\right)^2\right)}{2\sigma^2}\right), x_j \in U$$

$$h_{ji} = \exp\left(-\frac{\left\|\boldsymbol{\omega}^T[\boldsymbol{I}\otimes\Phi(x_j)]-\boldsymbol{\omega}^T[\boldsymbol{I}\otimes\Phi(x_i)]\right\|_2^2}{2\sigma^2}\right), x_i, x_j \in U$$

$$v_{ji} = \exp\left(-\frac{\left\|\boldsymbol{\omega}^T[\boldsymbol{I}\otimes\Phi(x_j)]-y_i\right\|_2^2}{2\sigma^2}\right), x_j \in U, y_i \in y$$

The objective function (13) can be solved by the alternating optimization strategy. Firstly, we fix $\boldsymbol{\alpha}$, and the objective function is to find the optimal classifier $\boldsymbol{\omega}$. It can be solved by the alternating direction method of multipliers (ADMM) [4,24]. Secondly, we fix $\boldsymbol{\omega}$ that is obtained in the first step, the objective function becomes

$$\arg\max_{\boldsymbol{\alpha}^T 1=1,\alpha_i\in\{0,1\}} \boldsymbol{\alpha}^T a + \beta_1\boldsymbol{\alpha}^T b - \beta_2\boldsymbol{\alpha}^T c \tag{14}$$

$$a_j = \sum_{k=1}^{C} \exp\left(-\frac{\left(1 + 2\left|\omega_k^T \Phi(x_j)\right| + \left(\omega_k^T \Phi(x_j)\right)^2\right)}{2\sigma^2}\right)$$

$$where\, b_j = \frac{1}{u} \sum_{x_i \in U} \exp\left(-\frac{\left\|\boldsymbol{\omega}^T[\boldsymbol{I} \otimes \Phi(x_j)] - \boldsymbol{\omega}^T[\boldsymbol{I} \otimes \Phi(x_i)]\right\|_2^2}{2\sigma^2}\right) w_{ji}$$

$$c_j = \frac{1}{l} \sum_{x_i \in L} \exp\left(-\frac{\left\|\boldsymbol{\omega}^T[\boldsymbol{I} \otimes \Phi(x_i)] - y_i\right\|_2^2}{2\sigma^2}\right) w_{ji}$$

To solve (14), as in [5], we relax α_j to a continuous range [0, 1]. Thus, the $\boldsymbol{\alpha}$ can be solved with a linear program. The sample corresponding to the largest value in α will be queried as x_q. The RMLAL algorithm is summarized in Algorithm 1.

Algorithm 1. Robust Multi-label Active Learning

Input: Labeled data set L and unlabeled data set U, the tradeoff parameters β_1 and β_2, and initial variables and parameters.

1: **repeat**
2: Fixed $\boldsymbol{\alpha}$, calculate the function (13) with ADMM strategy to obtain the values of ω in kernel space with $\omega_k = \sum_{x_i \in L} \theta_{ki} \Phi(x_i)$, where $\theta_k = [\theta_{k1}, \theta_{k2}, \ldots, \theta_{kl}]^T$ are auxiliary variables.
3: With the values of ω, calculate the indicator vector $\boldsymbol{\alpha}$ by solving(14), and select the sample that is corresponding to the largest value in $\boldsymbol{\alpha}$.
4: **until** the tolerance is satisfied

Output: The query index of unlabeled samples.

4 Experiments

4.1 Settings

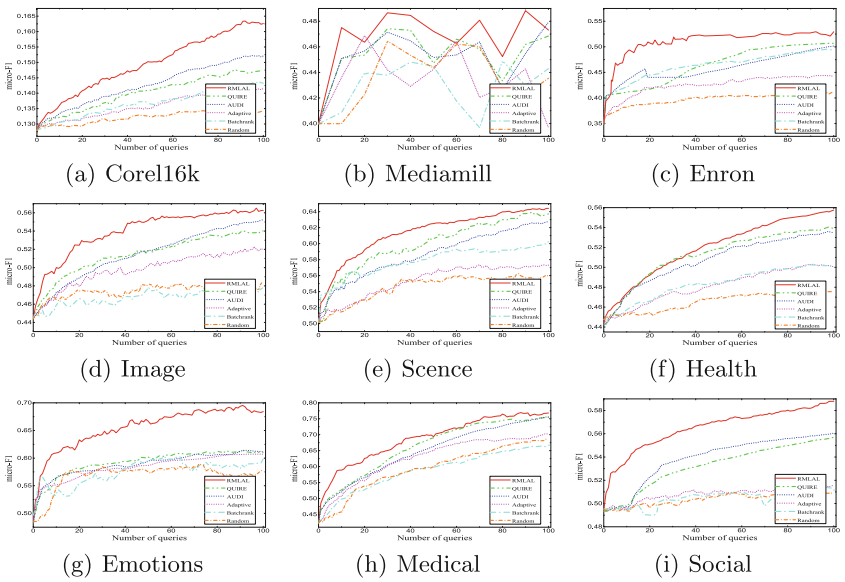
In this section, we present the experimental results to validate the effectiveness of the proposed method that compares with the prior methods on 9 multi-label data sets from Mulan project[1]. The characteristics of data sets are described in Table 1. To demonstrate the superior of our method, several methods are listed as follows as competitors.

1. RANDOM is the baseline which randomly selects instance label pairs.
2. AUDI [16] combines label ranking with threshold learning, then exploits both uncertainty and diversity in the instance space as well as the label space.
3. Adaptive [20] combines the max-margin prediction uncertainty and the label cardinality inconsistency as the criterion for active selection.

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

4. QUIRE [15] provides a systematic way for measuring and combining the informativeness and representativeness of an unlabeled instance by incorporating the correlation among labels.
5. Batchrank [5] selects the best query with an NP-hard optimization problem based on the mutual information.
6. RMLAL: Robust Multi-label Active Learning is the proposed in this paper.

LC is the average number of label for each instance in the data set. For each data set, we randomly divide it into two equal parts. One is regarded as the testing data set. For the other part, we randomly select 4 % samples as the initial labeled set, and the remaining samples of this part are used as the unlabeled data set for active learning. In the compared methods, AUDI and QUIRE query the relevance of an instance-label pairs in each iteration. We can notice that querying all labels for one instance is equal to query C label-instance pairs. Hence, for fair comparison, we query C label-instance pairs as one query instance in AUDI and QUIRE. For the method Batchrank, in the original paper, the tradeoff parameter sets as 1. For a fair comparison, we choose the tradeoff parameter from a candidate set that is the same in the proposed method. The parameters of other methods are all set as the same in original papers. For the kernel parameters, we adopt the same value for all methods.



(a) Corel16k (b) Mediamill (c) Enron

(d) Image (e) Scence (f) Health

(g) Emotions (h) Medical (i) Social

Fig. 4. Comparison of different active learning methods on fifteen benchmark datasets. The curves show the micro-F1 accuracy over queries, and each curve represents the average result of 5 runs.

Without loss of generality, the liblinear[2] is adopted as the classifier for all methods [9], and micro-F1 is used to evaluate the performance [5], which is a commonly used performance measurement in multi-label learning. For each data set, we repeat each method for 5 times and report the average results. The querying process stops when 100 iterations are reached and one instance is queried at each iteration.

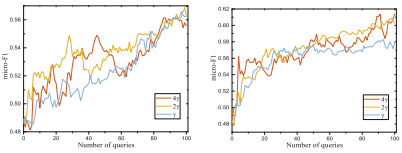
4.2 Results

For each data set, we reported the average results in Fig. 4. From all these results, we could observe that the proposed method performs the best on most of the data sets. It achieved the best results in almost the whole active learning process. In general, QUIRE and AUDI were two methods to query the label-instance pairs for labeling. They almost showed the superior performance to the Batchrank and Adaptive, which queried all labels for the instance. This demonstrated that querying the relevant labels was more efficient than querying all labels for one instance. But for our methods, it achieved the best performance with querying all labels for one instance than querying the relevant label-instance pairs. The reason may be that although the Batchrank and Adaptive queried all the labels, they could not avoid the influence of the outlier labels, leading to the query samples undesirable. For QUIRE and AUDI methods, some labels information lost when they just queried the limited relevant labels, and they need much samples to achieve a better performance. The results demonstrated the proposed method not only could achieve discriminative labeling but also could avoid the influence of the outlier labels. To put in nutshell, the proposed method merging the uncertainty and representativeness with MCC can solve the problems in multi-label active learning effectively as stated above.

4.3 Evaluation Parameters

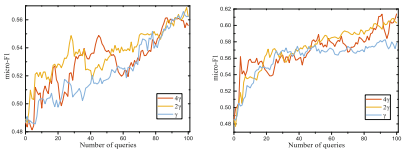
In the proposed method, the kernel parameter σ is very important for the MCC. There are two tradeoff parameters on the uncertain part and representative part respectively. For conveniently, in our experiments, we defined kernel size $\gamma = 1/(2 * \sigma^2)$, and we fixed the kernel size as $1/C$ in the label space. For the feature space, we fixed the kernel size as $1/dim$ in feature space, where dim is the dimension of feature space. To discover the influence of the kernel size for the proposed method, we evaluated the kernel size for MCC in label space. We reported the average results when the kernel size was set as $\{\gamma, 2\gamma, 4\gamma\}$ respectively on two popular benchmark datasets emotions and scence [5], which had the same number of labels but with different LC. For the tradeoff parameters, we chose them from a fixed candidate set $\{1, 10, 100\}$ respectively, and we also reported the average results on the two data sets. The other settings were same to the previous experiments. Figure 5 showed the average results with the kernel size changing. We can observe that the results are not very sensitive to the kernel size. This may be that

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



(a) Scence (b) Emotions

Fig. 5. Comparison of different γ on two data sets



(a) Scence (b) Emotions

Fig. 6. Comparison of different tradeoff parameter pairs (β_1, β_2) on two data sets

the changing of the parameter γ just changes the relative value of the discriminative labels and outlier labels with MCC, but the value of discriminative labels with MCC are always larger than that of outlier labels. Relatively, we can set the kernel size as double of γ for a better selection. Figure 6 showed the results with different pairs of the tradeoff parameters. For these results, we can observe that uncertain information and representative information have a big influence on the results. However, the better results are obtained in contrast on the two data sets. The scence data obtains the good results when β_1 is small and β_2 is large, while the emotions data obtained the good results when β_1 is large and β_2 is small. This may be that the LC of scence is small than the emotions data, leading to the initial labeled information of scence is less. With so little supervised information, the labeled data become important to build a query model. When LC is large, the supervised information may be redundant, and the unlabeled data become important. Therefore, the tradeoff parameters can be adopted according to the different data sets adaptively with LC.

5 Conclusion

Outlier labels are very common in multi-label scenarios and may cause the supervised information bias. In this paper, we propose a robust multi-label active learning based on MCC to solve the problem. The proposed method queries the samples that can not only build training models with a good generalization ability but also represent the similarity well for multi-label data. With MCC, the supervised information of outlier labels will be suppressed, and that of discriminative labels will be expanded. It outperformed state-of-the-art methods in most of the experiments. The experimental analysis also reveals that it is beneficial to update the trade-off parameter that balances the uncertain and representative information during the query process. We plan to develop an adaptive mechanism to tune this parameter automatically to make our algorithm more practical.

Acknowledgements. This work was supported in part by the National Basic Research Program of China (973 Program) under Grant 2012CB719905, the National Natural Science Foundation of China under Grants 61471274, 41431175, 61401317, U1536204, 60473023, 61302111, and the Australian Research Council Projects DP-140102164, FT-130101457, and LE140100061.

References

1. Agrawal, P., Girshick, R., Malik, J.: Analyzing the performance of multilayer neural networks for object recognition. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8695, pp. 329–344. Springer, Heidelberg (2014). doi:10.1007/978-3-319-10584-0_22
2. Bezdek, J.C., Hathaway, R.J.: Convergence of alternating optimization. Neural Parallel Sci. Comput. **11**(4), 351–368 (2003)
3. Boyd, S., Vandenberghe, L.: Convex optimization. Cambridge University Press, Cambridge (2004)
4. Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J.: Distributed optimization and statistical learning via the alternating direction method of multipliers. Found. Trends Mach. Learn. **3**(1), 1–122 (2011)
5. Chakraborty, S., Balasubramanian, V., Sun, Q., Panchanathan, S., Ye, J.: Active batch selection via convex relaxations with guaranteed solution bounds. TPAMI **37**(10), 1945–1958 (2015)
6. Chen, X., Shrivastava, A., Gupta, A.: Neil: extracting visual knowledge from web data. In: CVPR, pp. 1409–1416 (2013)
7. Chen, Y., Krause, A.: Near-optimal batch mode active learning and adaptive submodular optimization. In: CVPR, pp. 160–168 (2013)
8. Elhamifar, E., Sapiro, G., Yang, A., Sasrty, S.: A convex optimization framework for active learning. In: ICCV, pp. 209–216 (2013)
9. Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: Liblinear: a library for large linear classification. JMLR **9**, 1871–1874 (2008)
10. Freytag, A., Rodner, E., Denzler, J.: Selecting influential examples: active learning with expected model output changes. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8692, pp. 562–577. Springer, Heidelberg (2014). doi:10.1007/978-3-319-10593-2_37
11. He, R., Tan, T., Wang, L., Zheng, W.S.: $l_{2,1}$ regularized correntropy for robust feature selection. In: CVPR, pp. 2504–2511. IEEE (2012)
12. He, R., Zheng, W.S., Hu, B.G.: Maximum correntropy criterion for robust face recognition. CVPR **33**(8), 1561–1576 (2011)
13. He, R., Zheng, W.S., Tan, T., Sun, Z.: Half-quadratic-based iterative minimization for robust sparse representation. TPAMI **36**(2), 261–275 (2014)
14. Huang, S.J., Chen, S., Zhou, Z.H.: Multi-label active learning: query type matters. In: IJCAI, pp. 946–952. AAAI Press (2015)
15. Huang, S.J., Jin, R., Zhou, Z.H.: Active learning by querying informative and representative examples. TPAMI **36**(10), 1936–1949 (2014)
16. Huang, S.J., Zhou, Z.H.: Active query driven by uncertainty and diversity for incremental multi-label learning. In: ICDM, pp. 1079–1084. IEEE (2013)
17. Jing, L., Yang, L., Yu, J., Ng, M.K.: Semi-supervised low-rank mapping learning for multi-label classification. In: CVPR, June 2015
18. Kading, C., Freytag, A., Rodner, E., Bodesheim, P., Denzler, J.: Active learning and discovery of object categories in the presence of unnameable instances. In: CVPR, pp. 4343–4352. IEEE (2015)
19. Li, X.X., Dai, D.Q., Zhang, X.F., Ren, C.X.: Structured sparse error coding for face recognition with occlusion. TIP **22**(5), 1889–1900 (2013)
20. Li, X., Guo, Y.: Active learning with multi-label SVM classification. In: IJCAI. Citeseer (2013)

21. Li, X., Guo, Y.: Multi-level adaptive active learning for scene classification. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8695, pp. 234–249. Springer, Heidelberg (2014). doi:10.1007/978-3-319-10584-0_16
22. Liu, C., Yuen, J., Torralba, A., Sivic, J., Freeman, W.T.: SIFT flow: dense correspondence across different scenes. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008. LNCS, vol. 5304, pp. 28–42. Springer, Heidelberg (2008). doi:10.1007/978-3-540-88690-7_3
23. Liu, T., Tao, D.: Classification with noisy labels by importance reweighting. TPAMI **38**(3), 447–461 (2016)
24. Liu, T., Tao, D., Song, M., Maybank, S.J.: Algorithm-dependent generalization bounds for multi-task learning. TPAMI (2016). doi:10.1109/TPAMI.2016.2544314
25. Long, C., Hua, G.: Multi-class multi-annotator active learning with robust gaussian process for visual recognition. In: ICCV, December 2015
26. Mac Aodha, O., Campbell, N., Kautz, J., Brostow, G.: Hierarchical subquery evaluation for active learning on a graph. In: CVPR, pp. 564–571 (2014)
27. Qian, J., Yang, J., Zhang, F., Lin, Z.: Robust low-rank regularized regression for face recognition with occlusion. In: CVPRW, pp. 21–26 (2014)
28. Settles, B.: Active learning literature survey. University of Wisconsin, Madison, vol. 52, no. 55–66, p. 11 (2010)
29. Settles, B.: Active learning. Synth. Lect. Artif. Intell. Mach. Learn. **6**(1), 1–114 (2012)
30. Singh, G., Kosecka, J.: Nonparametric scene parsing with adaptive feature relevance and semantic context. In: CVPR, pp. 3151–3157 (2013)
31. Singh, M., Curran, E., Cunningham, P.: Active learning for multi-label image annotation. In: ICAIC, pp. 173–182 (2009)
32. Tao, D., Li, X., Xindong, W., Maybank, S.: General tensor discriminant analysis and gabor features for gait recognition. TPAMI **29**(10), 1700–1715 (2007)
33. Tao, D., Li, X., Xindong, W., Maybank, S.: Geometric mean for subspace selection. TPAMI **31**(2), 260–274 (2009)
34. Tao, D., Tang, X., Li, X., Wu, X.: Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval. TPAMI **28**(7), 1088–1099 (2006)
35. Vijayanarasimhan, S., Grauman, K.: Large-scale live active learning: training object detectors with crawled data and crowds. IJCV **108**(1–2), 97–114 (2014)
36. Xiong, X., Torre, F.: Supervised descent method and its applications to face alignment. In: CVPR, pp. 532–539 (2013)
37. Xu, C., Tao, D., Xu, C.: Multi-view intact space learning. TPAMI **37**(12), 2531–2544 (2015)
38. Yan, R., Yang, J., Hauptmann, A.: Automatically labeling video data using multi-class active learning. In: CVPR, pp. 516–523. IEEE (2003)
39. Yang, Y., Ma, Z., Nie, F., Chang, X., Hauptmann, A.G.: Multi-class active learning by uncertainty sampling with diversity maximization. IJCV **113**(2), 113–127 (2015)
40. Yuan, X.T., Hu, B.G.: Robust feature extraction via information theoretic learning. In: ICML. ACM (2009)
41. Zha, Z.J., Wang, M., Zheng, Y.T., Yang, Y., Hong, R., Chua, T.S.: Interactive video indexing with statistical active learning. TMM **14**(1), 17–27 (2012)
42. Zhao, F., Huang, Y., Wang, L., Tan, T.: Deep semantic ranking based hashing for multi-label image retrieval. In: CVPR, June 2015