Querying Discriminative and Representative Samples for Batch Mode Active Learning

Zheng Wang
Arizona State University
Tempe, AZ 85287, USA
zhengwang@asu.edu

Jieping Ye
Arizona State University
Tempe, AZ 85287, USA
jieping.ye@asu.edu

ABSTRACT

Empirical risk minimization (ERM) provides a useful guideline for many machine learning and data mining algorithms. Under the ERM principle, one minimizes an upper bound of the true risk, which is approximated by the summation of empirical risk and the complexity of the candidate classifier class. To guarantee a satisfactory learning performance, ERM requires that the training data are i.i.d. sampled from the unknown source distribution. However, this may not be the case in active learning, where one selects the most informative samples to label and these data may not follow the source distribution. In this paper, we generalize the empirical risk minimization principle to the active learning setting. We derive a novel form of upper bound for the true risk in the active learning setting; by minimizing this upper bound we develop a practical batch mode active learning method. The proposed formulation involves a non-convex integer programming optimization problem. We solve it efficiently by an alternating optimization method. Our method is shown to query the most informative samples while preserving the source distribution as much as possible, thus identifying the most uncertain and representative queries. Experiments on benchmark data sets and real-world applications demonstrate the superior performance of our proposed method in comparison with the state-of-the-art methods.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Data Mining*

General Terms

Algorithms

Keywords

Active learning, representative and discriminative, empirical risk minimization, maximum mean discrepancy

KDD'13, August 11–14, 2013, Chicago, Illinois, USA.

1. INTRODUCTION

In many machine learning tasks, we need to collect the training data and manually annotate them by experts. This procedure is very expensive in most real world applications, such as text classification [34], collaborative filtering [23], outlier detection [1], biomedicine and bioinformatics [33]. Active learning is a very useful tool in such situations when unlabeled data is cheap to collect but labeling them is expensive. There are two main intuitions for querying the unlabeled samples and designing practical active learning algorithms [14]. The first one is to find the most informative or discriminative samples for the current classifier. This mechanism will shrink the space of candidate classifiers as rapidly as possible. The most typical criteria of this kind include expected error reduction [25], query by committee [17, 27] and the most uncertain rule [8, 26, 30]. In such methods, the queried data are not guaranteed to be i.i.d. sampled from the original data distribution, as they are selectively sampled based on the active learning criterion [3]. When training the classifier using the empirical risk minimization principle, this sampling bias prevents active learning from finding a classifier with good performance on future unseen data, and will also degrade the following query efficiency [14, 29]. The second category of active learning aims to alleviate this problem by querying the most representative samples for the overall patterns of the unlabeled data and preserving the data distribution or its statistics [11, 12, 35]. Such type of active learning methods gives better performance when there is few or no initial labeled data. However, their efficiency will degrade with the increase of queried labels, as they do not fully use the label information.

Since using either kind of criterion alone is not sufficient to get the optimal result, there are several works trying to query the unlabeled samples with both high informativeness and high representativeness [24, 34]. Usually they are either heuristic in designing the specific query criterion or ad hoc in measuring the informativeness and representativeness of the samples. Recently, Huang et al. [22] try to use both discriminative and representative information in one optimization formulation. They use the most uncertainty as the query criterion, and use unlabeled data in the semi-supervised learning setting for boosting the learning performance. However, the queried samples may not preserve the original data distribution. If the data structure does not satisfy the semi-supervised assumptions [10, 36], they may not achieve good performance.

In this paper, we extend the empirical risk minimization principle to the active learning case and present a novel ac-

tive learning framework. In this framework, we adapt maximum mean discrepancy (MMD) [5, 18, 28] to measure the distribution difference and derive an empirical upper bound for active learning risk. By minimizing this upper bound, we approximately minimize the true risk under the original data distribution. We propose a practical batch mode active learning algorithm under this framework. In our algorithm, we seek to query a subset of unlabeled samples which help minimize the generalization risk, based on all available information. To achieve this goal, the samples we query not only help to rapidly reduce the empirical risk on the training data, but also preserve the original data distribution, resulting in a good generalization ability for the unseen samples. This leads to a proper use of both discriminative information and representative information simultaneously. Moreover, using our active learning method, we can naturally handle the situations with or without initial labeled samples and achieve high active learning efficiency in either case. We have conducted experiments on benchmark data sets and real-world applications. Results demonstrate the effectiveness of the proposed method in comparison with the state-of-the-art batch mode active learning methods.

The rest of this paper is organized as follows: Section 2 analyzes the empirical risk minimization principle in the active learning setting and presents the corresponding active learning framework; in Section 3 we propose a practical batch mode active learning algorithm under our novel framework; experimental results are reported in Section 4; Section 5 concludes this paper and discusses the future work.

2. EMPIRICAL RISK MINIMIZATION FOR ACTIVE LEARNING

In supervised learning, the target of learning is to find the optimal classifier which is expected to generalize well on the unseen data. The empirical risk minimization (ERM) is a successful guideline for designing machine learning and data mining methods [7, 31]. It minimizes an upper bound of the true risk under the unknown data distribution. This upper bound is approximated by the summation of empirical risk on the available data and a properly designed regularization term, which constrains the complexity of the candidate classifiers [31, 2]. Assume we are given a data source D, with unknown distribution $p(\mathbf{z}) = p(\mathbf{x}, y)$ for sample $\mathbf{z} = \{\mathbf{x}, y\}$, and a finite data set S with n points, which are i.i.d. sampled from the same distribution, $p(\mathbf{z})$. Using the Rademacher complexity to describe the complexity of the function class, we obtain the uniform convergence property between the true risk and the empirical risk [2]:

$$\mathrm{E}_D(l(\mathbf{z})) \leq \hat{\mathrm{E}}_S(l(\mathbf{z})) + 2R_n(\mathcal{L}) + \sqrt{\frac{\ln 1/\delta}{n}}, \qquad (1)$$

which holds with probability at least $1-\delta$. In this inequality, $l(\mathbf{z}) \in \mathcal{L}$ is the loss function and $l(\mathbf{z}) = l(f(\mathbf{x}), y)$ for the classifier $f(\mathbf{x}) \in \mathcal{F}$. The true risk is defined as the expectation of the loss function:

$$\mathrm{E}_D(l(\mathbf{z})) = \int_{\mathbf{z}\in D} l(\mathbf{z}) d\mathbf{z}. \qquad (2)$$

The empirical risk is the empirical average of the loss function:

$$\hat{\mathrm{E}}_S(l(\mathbf{z})) = \frac{1}{|S|}\sum_{z\in S} l(\mathbf{z}). \qquad (3)$$

The Rademacher complexity of the loss function class $\mathcal{L}$ is expressed as

$$R_n(\mathcal{L}) = \mathrm{E}_S\Big[\mathrm{E}_\sigma\Big[\sup_{l\in\mathcal{L}}\big(\frac{1}{n}\sum_{i=1}^{n}\sigma_i l(\mathbf{z}_i)\big)\Big]\Big],$$

where $\sigma_1, \cdots, \sigma_n$ are independent random variables uniformly chosen from $\{-1, 1\}$, known as Rademacher variables.

In this framework, the empirical average (3) is under the same sample distribution as the expectation (2). This requires data in S to be i.i.d. sampled from the original data distribution $p(\mathbf{x}, y)$. However, this assumption may not hold in the active learning setting. In active learning, we assume that the labeled data are selectively sampled from another data distribution $q(\mathbf{x}, y)$, which is usually different from the distribution $p(\mathbf{x}, y)$ for the original problem. To extend the ERM principle to active learning, we reformulate the risk bound inequality as:

$$\begin{aligned}\mathrm{E}_D(l(\mathbf{z})) \leq\ & (\mathrm{E}_D(l(\mathbf{z})) - \mathrm{E}_Q(l(\mathbf{z}))) \\ & + \hat{\mathrm{E}}_Q(l(\mathbf{z})) + 2R_q(\mathcal{L}) + \sqrt{\tfrac{\ln 1/\delta}{q}}.\end{aligned} \qquad (4)$$

$\hat{\mathrm{E}}_Q(l(\mathbf{z}))$ is the empirical risk for the available labeled data, which may include initial labeled samples and query samples. $R_q(\mathcal{L})$ is the Rademacher complexity based on these labeled samples. There is a new term in the upper bound, which is the difference between the true risk under different data distributions:

$$\mathrm{E}_D\left(l(\mathbf{z})\right) - \mathrm{E}_Q(l(\mathbf{z}))\,.$$

Though in active learning the data distribution for the labeled samples $q(\mathbf{x}, y)$ may be different from the original distribution $p(\mathbf{x}, y)$, they share the same conditional probability $p(y|\mathbf{x})$. Let $p(\mathbf{x}, y) = p(\mathbf{x})p(y|\mathbf{x})$ and $q(\mathbf{x}, y) = q(\mathbf{x})p(y|\mathbf{x})$, we rewrite the first term in the upper bound of (4) as

$$\begin{aligned}& \mathrm{E}_D(l(z)) - \mathrm{E}_Q(l(z)) \\ =\ & \textstyle\int_{\mathbf{x}} p(\mathbf{x}) \int_y l(f(\mathbf{x}), y) p(y|\mathbf{x}) dy d\mathbf{x} \\ & - \textstyle\int_{\mathbf{x}} q(\mathbf{x}) \int_y l(f(\mathbf{x}), y) p(y|\mathbf{x}) dy d\mathbf{x} \\ =\ & \textstyle\int_{\mathbf{x}} g(\mathbf{x}) p(\mathbf{x}) d\mathbf{x} - \int_{\mathbf{x}} g(\mathbf{x}) q(\mathbf{x}) d\mathbf{x},\end{aligned}$$

where we define $g(\mathbf{x}) = \int_y\ l(f(\mathbf{x}), y)\ p(y|\mathbf{x})\ dy$. In learning problems, the prediction functions have bounded norm $||f||_{\mathcal{F}}$. Thus, given a continues loss function, such as the hinge loss and the least squares loss [31], the function g is bounded. Since g is also measurable, there exists a bounded and continuous function $\hat{g}$ which has the following property [15]:

$$\begin{aligned}& \textstyle\int_{\mathbf{x}} g(\mathbf{x}) p(\mathbf{x}) d\mathbf{x} - \int_{\mathbf{x}} g(\mathbf{x}) q(\mathbf{x}) d\mathbf{x} \\ =\ & \textstyle\int_{\mathbf{x}} \hat{g}(\mathbf{x}) p(\mathbf{x}) d\mathbf{x} - \int_{\mathbf{x}} \hat{g}(\mathbf{x}) q(\mathbf{x}) d\mathbf{x} \\ \leq\ & \textstyle\sup_{\hat{g}\in\mathcal{C}(\mathbf{x})} \int_{\mathbf{x}} \hat{g}(\mathbf{x}) p(\mathbf{x}) d\mathbf{x} - \int_{\mathbf{x}} \hat{g}(\mathbf{x}) q(\mathbf{x}) d\mathbf{x},\end{aligned}$$

where $\hat{g}$ belongs to the function class of bounded and continuous functions $\mathcal{C}(\mathbf{x})$ of $\mathbf{x}$. From [5, 18, 28], we find that the right side of the inequality is the maximum mean discrepancy term defined as

$$\mathrm{MMD}[\mathcal{C}, p(\mathbf{x}), q(\mathbf{x})] = \sup_{\hat{g}\in\mathcal{C}(\mathbf{x})} \int_{\mathbf{x}} \hat{g}(\mathbf{x}) p(\mathbf{x}) d\mathbf{x} - \int_{\mathbf{x}} \hat{g}(\mathbf{x}) q(\mathbf{x}) d\mathbf{x}.$$

Taking MMD as an upper bound of the expected risk difference, the ERM risk bound for active learning can be written as

$$\mathrm{E}_D(l(f(\mathbf{x}),y)) \leq \hat{\mathrm{E}}_Q(l(f(\mathbf{x}),y)) + \mathrm{MMD}\left[\mathcal{C}, p(\mathbf{x}), q(\mathbf{x})\right] + \left[2R_q(\mathcal{L}) + \sqrt{\tfrac{\ln(1/\delta)}{q}}\right].$$

Following [5, 18, 28], we could empirically restrict the MMD on a reproducing kernel Hilbert space (RKHS) with a characteristic kernel, $k(\mathbf{x}_i, \mathbf{x}_j)$, which is associated with a nonlinear feature mapping function $\phi(\mathbf{x})$. Then the ERM principle in the active learning case is summarized in the following theorem. The proof is provided in the Appendix.

THEOREM 2.1. *Assume that the kernel function is upper bounded by a constant, $0 \leq k(\mathbf{x}_i, \mathbf{x}_j) \leq M$, $\forall\ \mathbf{x}_j, \mathbf{x}_j$. Let the variables be defined as above. Under the ERM principle for active learning, the following holds with probability at least $1-\delta$,*

$$\mathrm{E}_D(l(f(\mathbf{x}),y)) \leq \hat{\mathrm{E}}_Q(l(f(\mathbf{x}),y)) + \mathrm{MMD}_\phi(S,Q) + C(\mathcal{L},q,\delta). \tag{5}$$

In this inequality, the empirical MMD term is

$$\mathrm{MMD}_\phi(S,Q) = \left\| \frac{1}{n}\sum_{\mathbf{x}_i \in S} \phi(\mathbf{x}_i) - \frac{1}{q}\sum_{\mathbf{x}_i \in Q} \phi(\mathbf{x}_i) \right\|_{\mathcal{F}}.$$

The function class complexity term is

$$C(\mathcal{L},q,\delta) = 2R_q(\mathcal{L}) + c\sqrt{\frac{M\ln(1/\delta)}{q}},$$

where c is a constant.

3. BATCH MODE DISCRIMINATIVE AND REPRESENTATIVE ACTIVE LEARNING

Suppose we are given a data set with n samples $S = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$ of d dimensions. Initially we have l labeled samples. Without loss of generality, we denote them as $L = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_l, y_l)\}$, with labels $y_i \in \{-1, 1\}$, as we only focus on binary problems. Note that l could be 0. The remaining $u = n - l$ samples form the unlabeled set $U = \{\mathbf{x}_{l+1}, \mathbf{x}_{l+2}, \cdots, \mathbf{x}_n\}$, which is the candidate set for active learning. In our batch mode active learning problem, we iteratively select the best subset $Q \subset U$ with b samples to label, and put them to the labeled set L. In the following discussion, we use Q to denote the query sample set.

3.1 Active Learning with the ERM Principle

Based on Theorem 2.1, we propose a practical active learning algorithm by minimizing the active learning risk bound in (5). Mathematically, it is formulated as an optimization problem w.r.t. the classifier f and the query set Q:

$$\{Q^*, f^*\} = \arg\min_{Q,f} \sum_{\mathbf{x}\in L\cup Q} l(y, f(\mathbf{x})) + (l+b)\mathrm{MMD}_\phi(S, L\cup Q) + \lambda||f||^2_{\mathcal{F}}. \tag{6}$$

where $||f||^2_{\mathcal{F}}$ is used to constrain the complexity of the classifier class, which is equivalent to constraining $C(\mathcal{L}, b, \delta)$ [2]. $l(y, f(\mathbf{x}))$ in the objective function can be any popularly used loss function, such as the least squares loss, the hinge loss or the negative log likelihood of logistic regression. We choose the least squares loss for simplicity.

The optimization problem (6) is difficult to solve, as it involves a square root in the MMD term. Therefore, we substitute this term with its quadratic form, and obtain the following problem

$$\min_{Q,f} \sum_{\mathbf{x}_i\in L}(y_i - f(\mathbf{x}_i))^2 + \sum_{\mathbf{x}_i\in Q}(\hat{y}_i - f(\mathbf{x}_i))^2 + \lambda||f||^2_{\mathcal{F}} + \beta\mathrm{MMD}^2_\phi(S, L\cup Q)). \tag{7}$$

The optimal solution is not changed with a properly chosen parameter β [18, 31]. As we do not know the labels of the query samples before we get them manually labeled, we use the pseudo labels $\hat{y}_i$ in the objective, which are binary variables from $\{-1, 1\}$ [13]. In this objective function, the first three terms correspond to the regularized risk for all labeled samples after query, which carries the discriminative information embedded in the current classifier. We call them the discriminative part. The last term describes the distribution difference between the labeled samples after query and all available samples, which captures the representative information embedded in the labeled samples. The objective in (7) balances the discriminative and representative information in a single formulation. In the remaining part of this section, we will analyze this objective in a specific form and propose a practical batch mode active learning algorithm to solve the resulting optimization problem.

3.2 Discriminative Information by the Uncertainty of Minimum Margin

First, we show how to determine the b unknown pseudo labels $\hat{y}_i$. It is clear that the maximum possible regularized empirical risk after querying the b samples in Q is

$$\max_{\hat{y}_i:\forall \mathbf{x}_i\in Q} \min_{Q,f} \sum_{\mathbf{x}_i\in L}(y_i - f(\mathbf{x}_i))^2 + \sum_{\mathbf{x}_i\in Q}(\hat{y}_i - f(\mathbf{x}_i))^2 + \lambda||f||^2_{\mathcal{F}}. \tag{8}$$

If we solve (8) w.r.t. $\hat{y}_i$ with fixed Q and f, we minimize the worst-case risk introduced by the query samples. In this case, the pseudo labels are given by $\hat{y}_j = -\mathrm{sign}(f(\mathbf{x}_j))$. Accordingly, the related risk terms become

$$\min_{Q,f} \sum_{\mathbf{x}_i\in L}(y_i - f(\mathbf{x}_i))^2 + \sum_{\mathbf{x}_i\in Q}\Big[f(\mathbf{x}_i)^2 + 2|f(\mathbf{x}_i)| + 1\Big] + \lambda||f||^2_{\mathcal{F}}, \tag{9}$$

which is still an upper bound of the true risk. For any classifier f, (9) identifies the samples with minimum margin summation, given by

$$\min_{Q} \sum_{\mathbf{x}_i\in Q} |f(\mathbf{x}_i)|.$$

Intuitively, it looks for the most uncertain query samples.

We use the linear regression model in the kernel space as the classifier, which is in the form of $f(\mathbf{x}) = \mathbf{w}^T\phi(\mathbf{x})$, with the feature mapping $\phi(\mathbf{x})$. The discriminative part of our objective becomes

$$\min_{Q,\mathbf{w}} \sum_{\mathbf{x}_i\in L}(y_i - \mathbf{w}^T\phi(\mathbf{x}_i))^2 + \lambda||\mathbf{w}||^2 + \sum_{\mathbf{x}_i\in Q}\Big[(\mathbf{w}^T\phi(\mathbf{x}_i))^2 + 2|\mathbf{w}^T\phi(\mathbf{x}_i)|\Big].$$

3.3 Representative Information by Distribution Matching of MMD

The representative part in objective (7) is the MMD term, which is used to constrain the distribution of the labeled and query samples, and make it similar to the overall sample distribution as much as possible. It captures the representative information of the data structure. This part is empirically calculated as [5, 18, 28]:

$$\mathrm{MMD}^2_\phi(D, L\cup Q) = \left\| \frac{1}{n}\sum_{\mathbf{x}_i\in S}\phi(\mathbf{x}_i) - \frac{1}{l+b}\sum_{\mathbf{x}_i\in L\cup Q}\phi(\mathbf{x}_i)\right\|^2_{\mathcal{F}}.$$

Similar to [11], we transfer the MMD term into

$$\frac{1}{2}\boldsymbol{\alpha}^T K_{UU}\boldsymbol{\alpha} + \frac{u-b}{n}\mathbf{1}_l K_{LU}\boldsymbol{\alpha} - \frac{l+b}{n}\mathbf{1}_u K_{UU}\boldsymbol{\alpha} + \text{constant},$$

where $\mathbf{1}_l$ is a vector of length l, with all entries 1; $\mathbf{1}_u$ is of length u; $\boldsymbol{\alpha}$ is the indicator vector with u elements and each element $\alpha_i \in \{0,1\}$, and $\boldsymbol{\alpha}^T\mathbf{1}_u = b$. K is the kernel matrix with its element as $K_{ij} = k(\mathbf{x}_i,\mathbf{x}_j) = \phi(\mathbf{x}_i)^T\phi(\mathbf{x}_j)$, and K_{AB} denotes its sub-matrix between the samples from set A and set B. The objective can be further simplified as

$$\boldsymbol{\alpha}^T K_1\boldsymbol{\alpha} + \mathbf{k}\boldsymbol{\alpha}, \tag{10}$$

where $K_1 = \frac{1}{2}K_{UU}$, $\mathbf{k} = \mathbf{k}_3 - \mathbf{k}_2$, and $\forall\,\mathbf{x}_i \in U$, $\mathbf{k}_2(i) = \frac{l+b}{n}\sum_{\mathbf{x}_j\in U} K(i,j)$, $\mathbf{k}_3(i) = \frac{u-b}{n}\sum_{\mathbf{x}_j\in L} K(i,j)$.

3.4 The Proposed Formulation

Combining the discriminative and representative parts together, we obtain the following formulation:

$$\begin{aligned}\min_{\boldsymbol{\alpha}^T\mathbf{1}_u=b,\mathbf{w}} \quad & \textstyle\sum_{i=1}^l (y_i - \mathbf{w}^T\phi(\mathbf{x}_i))^2 + \lambda||\mathbf{w}||^2 \\ & + \textstyle\sum_{i=1}^u \alpha_i\Big[||\mathbf{w}^T\phi(\mathbf{x}_j)||_2^2 + 2|\mathbf{w}^T\phi(\mathbf{x}_j)|\Big] \\ & + \beta(\boldsymbol{\alpha}^T K_1\boldsymbol{\alpha} + \mathbf{k}\boldsymbol{\alpha}).\end{aligned} \tag{11}$$

This objective function approximates an upper bound of the generalization risk under the original data distribution. This problem is not convex, and we propose to employ the alternating optimization strategy [4].

If the query index $\boldsymbol{\alpha}$ is fixed, the objective is to find the best classifier based on the current labeled and query samples:

$$\begin{aligned}\min_{\mathbf{w}} \quad & \textstyle\sum_{i=1}^l (y_i - \mathbf{w}^T\phi(\mathbf{x}_i))^2 + \lambda||\mathbf{w}||^2 \\ & + \textstyle\sum_{j=1}^b \Big[||\mathbf{w}^T\phi(\mathbf{x}_j)||_2^2 + 2|\mathbf{w}^T\phi(\mathbf{x}_j)|\Big].\end{aligned} \tag{12}$$

We propose to solve (12) by the alternating direction method of multipliers (ADMM) [6].

If $\mathbf{w}$ is fixed, the objective becomes

$$\begin{aligned}\min_{\boldsymbol{\alpha}^T\mathbf{1}_u=b} \quad & \textstyle\sum_{i=1}^u \alpha_i\Big[(\mathbf{w}^T\phi(\mathbf{x}_i))^2 + 2|\mathbf{w}^T\phi(\mathbf{x}_i)|\Big] \\ & +\beta(\boldsymbol{\alpha}^T K_1\boldsymbol{\alpha} + \mathbf{k}\boldsymbol{\alpha}),\end{aligned} \tag{13}$$

which can be rewritten as

$$\min_{\boldsymbol{\alpha}^T\mathbf{1}_u=b} \beta\boldsymbol{\alpha}^T K_1\boldsymbol{\alpha} + (\beta\mathbf{k}+\mathbf{a})\boldsymbol{\alpha}, \tag{14}$$

where $a_j = ||\mathbf{w}^T\phi(\mathbf{x}_j)||_2^2 + 2|\mathbf{w}^T\phi(\mathbf{x}_j)|$. This is a quadratic programming problem for the indicator vector $\boldsymbol{\alpha}$. If we relax $\boldsymbol{\alpha}$ to continuous values in $[0,1]^u$, this can be solved using standard quadratic programming.

3.5 The Proposed Algorithm

We provide the details for solving the optimization problem (11), which is not convex. The alternating procedure includes two main steps: *step 1*: for a fixed $\boldsymbol{\alpha}$, employ the alternating direction method of multipliers (ADMM) to solve $\mathbf{w}$; *step 2*: for a fixed $\mathbf{w}$, employ the quadratic programming (QP) to solve $\boldsymbol{\alpha}$.

Step 1: Computing w, for a fixed $\boldsymbol{\alpha}$:

Using the kernel form, the problem is to learn $\boldsymbol{\tau}$ for $\mathbf{w} = \sum_{\mathbf{x}_j\in L}\tau_j\phi(\mathbf{x}_j)$ using the following formulation:

$$\begin{aligned}&\min_{\boldsymbol{\tau}} \textstyle\sum_{i=1}^l (y_i - \sum_{\mathbf{x}_j\in L}\tau_j K(\mathbf{x}_j,\mathbf{x}_i))^2 + \lambda\boldsymbol{\tau}^T K_{LL}\boldsymbol{\tau} \\ &+ \textstyle\sum_{i=1}^b \Big[||\sum_{\mathbf{x}_j\in L}\tau_j(\mathbf{x}_j,\mathbf{x}_i)||_2^2 + 2|\sum_{\mathbf{x}_j\in L}\tau_j K(\mathbf{x}_j,\mathbf{x}_i)|\Big].\end{aligned}$$

By introducing the auxiliary variable $z_j = \mathbf{w}^T\phi(\mathbf{x}_j)$, the objective function becomes,

$$\begin{aligned}\min_{\boldsymbol{\tau}} \quad & \sum_{i=1}^l (y_i - \boldsymbol{\tau}^T K_L(\mathbf{x}_i))^2 + \lambda\boldsymbol{\tau}^T K_{LL}\boldsymbol{\tau} + \sum_{i=1}^b \Big[z_i^2 + 2|z_i|\Big] \\ s.t. \quad & z_i - \boldsymbol{\tau}^T K_L(\mathbf{x}_i) = 0,\ \forall\,\mathbf{x}_i \in Q.\end{aligned} \tag{15}$$

We construct the augmented Lagrangian as

$$\begin{aligned}L_\rho \quad &= ||\mathbf{y}_L - \boldsymbol{\tau}^T K_{LL}||^2 + \lambda\boldsymbol{\tau}^T K_{LL}\boldsymbol{\tau} \\ &+||\mathbf{z}||^2 + 2|\mathbf{z}| + (\mathbf{z} - \boldsymbol{\tau}^T K_{LQ})\boldsymbol{\gamma}^T \\ &+(\rho/2)||\mathbf{z} - \boldsymbol{\tau}^T K_{LQ}||_2^2.\end{aligned}$$

Then we obtain the updating rules as

$$\begin{aligned}&\boldsymbol{\tau}^{k+1} = A^{-1}\mathbf{r}^T, \\ &\text{with } A = K_{LL}^2 + \frac{\rho}{2}K_{LQ}K_{QL} + \lambda K_{LL}, \\ &\text{and } \mathbf{r} = \mathbf{y}_L K_{LL} + \frac{1}{2}\boldsymbol{\gamma}^k K_{LQ}^T + \frac{\rho}{2}\mathbf{z}^k K_{LQ}^T; \\ &\mathbf{z}^{k+1} = \arg\min \frac{1}{2}||\mathbf{z} - \mathbf{v}||^2 + \eta|\mathbf{z}| = \text{sign}(\mathbf{v})(|\mathbf{v}| - \eta)_+, \\ &\text{with } \mathbf{v} = \frac{\rho(\boldsymbol{\tau}^{k+1})^T K_{LQ} - \boldsymbol{\gamma}^k}{\rho+2},\ \eta = \frac{2}{\rho+2}; \\ &\boldsymbol{\gamma}^{k+1} = \boldsymbol{\gamma}^k + \rho(\mathbf{z}^{k+1} - (\boldsymbol{\tau}^{k+1})^T K_{LQ}).\end{aligned} \tag{16}$$

Step 2: Computing $\boldsymbol{\alpha}$, for a fixed w:

With a fixed $\mathbf{w}$, the objective function becomes

$$\min_{\boldsymbol{\alpha}^T\mathbf{1}_u=b} \boldsymbol{\alpha}^T H\boldsymbol{\alpha} + \mathbf{d}\boldsymbol{\alpha}. \tag{17}$$

where $H = \beta K_1$ and $\mathbf{d} = \beta\mathbf{k}+\mathbf{a}$. This problem can be solved using standard QP toolboxes such as CVX[1] and MOSEK[2]. With the compute $\boldsymbol{\alpha}$, we set the largest b elements in $\boldsymbol{\alpha}$ to 1 and set the remaining ones to 0.

The key steps are summarized in Algorithm 1. We can also generalize our method to the semi-supervised setting, by introducing estimated empirical risk for all unlabeled samples as in [20, 22].

[1]CVX: "http://cvxr.com/cvx".
[2]MOSEK: "http://www.mosek.com/".

Algorithm 1 Discriminative and Representative Queries for Batch Mode Active Learning (BMDR)

Input: $L = \{(\mathbf{x}_i, y_i)\}$ with l labeled samples, $U = \{\mathbf{x}_i\}$ with u unlabeled samples, parameters λ, β, batch size b, tolerance ε for convergence condition
Initialize: Set initial variables and parameters.
repeat
Step 1: optimize the objective function (15) w.r.t $\boldsymbol{\tau}$ using ADMM, updated by (16).
Step 2: optimize the objective function (17) w.r.t $\boldsymbol{\alpha}$ using QP; set the largest b elements in $\boldsymbol{\alpha}$ to 1 and others to 0.
until Convergence condition is satisfied
Output: The query indicator vector $\boldsymbol{\alpha}$.

4. EXPERIMENTS

In our experiments, we compare our method with random selection and state-of-the-art batch mode active learning methods. We list all methods we compared in the experiments as follows:

1. *Random*: randomly select the query samples.
2. *Fbatch*: batch mode active learning based on fisher information [21].
3. *Dbatch*: discriminative batch mode active learning [20].
4. *Tbatch*: batch mode active learning using transductive experimental design [35].
5. *Mbatch*: batch mode active learning by matrix completion based on mutual information [19].
6. *BMDR*: our batch mode active leaning with discriminative and representative queries.

We conduct the experiments on fifteen data sets from UCI benchmarks[3] [16]: australian, banana, chess, crx, diabetis, heart, image, ionosphere, monk1, ringnorm, splice, thyroid, twonorm, vote and waveform. We summarize the characteristics of the data sets in Table 1.

In the experiments, for each data set, we use 60% data for training and 40% for testing, and the data set is randomly divided into training and test sets. We use the training set for active learning and compare the prediction accuracy for different methods on the test set. We assume there is no labeled data available at the very beginning of active learning. For Fbatch and Dbatch methods which need initial labeled data, we randomly sample the initial labeled data until there are enough labeled samples to train an initial classifier. The number of these initial samples are usually smaller than 10 in our experiments. The experiment stops when 80% of the training set has been labeled, or the learning accuracy does not increase for any method. This stopping criterion guarantees we show the whole active learning process, though practically the query process stops much earlier due to the limited labeling cost. We set the batch size $b = 5$ in all experiments. For the parameters involved in the competing methods, we prefer to use the values recommended in the original papers. In other cases, we set up a large candidate set and select the best parameter value. In our BMDR method, we set the regularization weight $\lambda = 0.1$, and the trade-off parameter β is chosen from a candidate set by cross validation. For each data set, we use the same kernel for all methods, which is properly chosen from the linear kernel or RBF kernel with the optimal kernel width. For fairness, we use the same SVM classifier for all methods to evaluate the informativeness of the selected samples. We report the accuracy curve of the SVM classifier after each query. We use the SVM implementation provided by LIBSVM [9]. In these experiments, we use the CVX toolbox as the solver for the quadratic programming problems and the linear programming problems. Running the Mbatch method needs a large amount of memory for large data sets. Though we could use subsampling to save the memory, it will degrade the performance of this method. For this reasons, we only provide the result for this method on relatively small data sets.

[3]Some of the data sets have been preprocessed and released at "http://theoval.cmp.uea.ac.uk/∼gcc/matlab/default.html#benchmarks".

Table 1: Characteristics of the data sets, including the numbers of the corresponding features and samples.

Data set	# Feature	# Instance
banana	2	4000
diabetis	8	768
heart	13	270
twonorm	20	7400
waveform	21	5000
ringnorm	20	7400
thyroid	5	215
chess	36	3196
ionosphere	34	351
splice	60	2991
vote	16	435
image	18	2086
crx	15	690
australian	14	690
monk1	6	432

4.1 Results

For each data set, we run the experiment independently for 10 times, and present the average result in Figure 1. We also show the significance of the comparison results using the paired t-test. In active learning, we need to compare the performance during the entire query process. In our experiments, we compare the learning accuracy of our method versus each competing method after each query, at 95% confidence level, then count the times of our win/tie/loss. We show them in percentage for all data sets in Table 2.

From all these results, we can observe that our method outperforms the competitors in three aspects. First, our method seldom performs worse than random query. All the other active learning methods are dominated by random query in certain cases. Second, the performance of our method is always among the best ones on all data sets. Third, in most cases, our method performs consistently better than the competitors during the whole active learning process. These results demonstrate that both discriminative and representative information are critical for active learn-

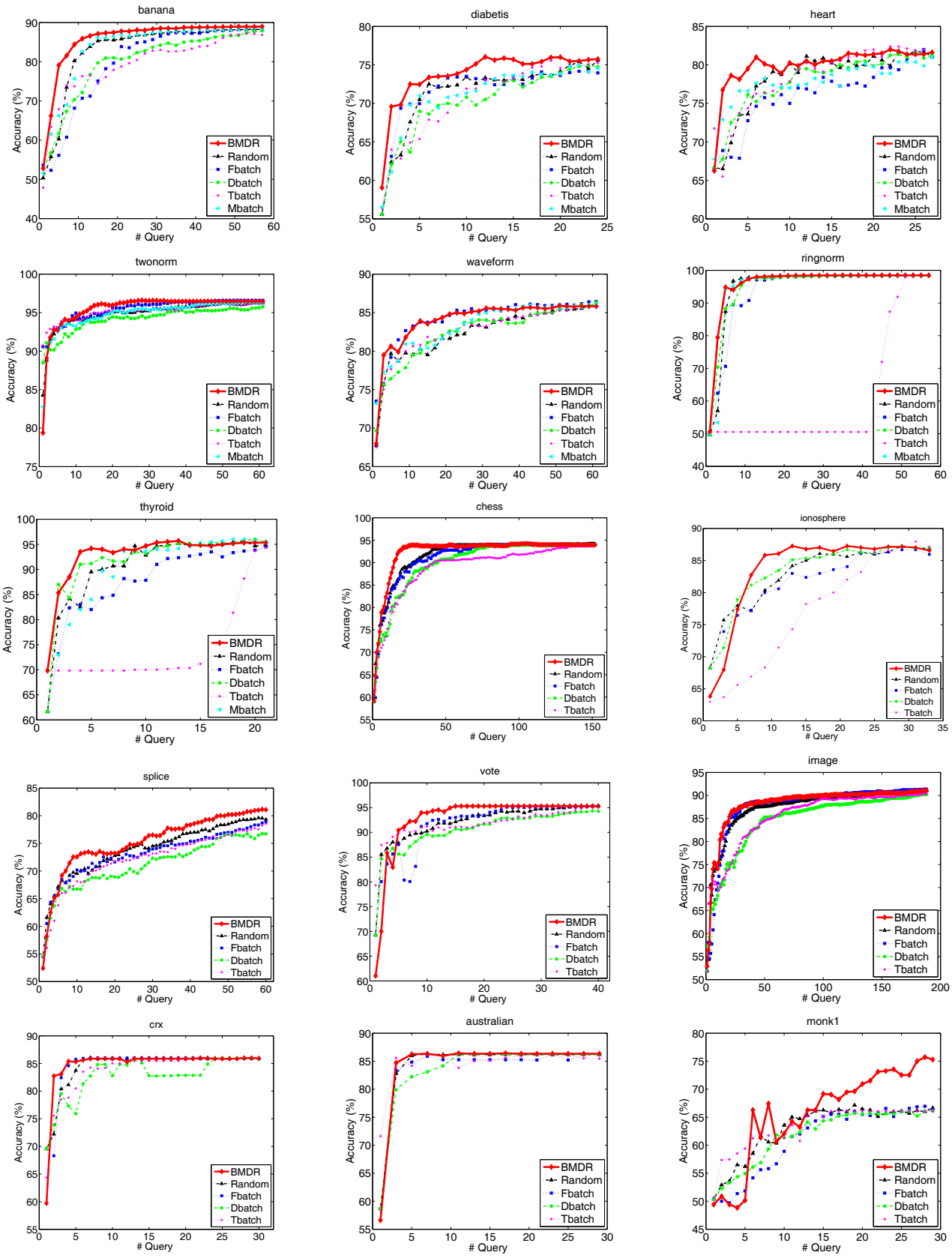

Figure 1: Comparison of different batch mode active learning methods on fifteen benchmark data sets. The curve shows the learning accuracy over queries, and each curve represents the average result of 10 runs.

Table 2: The win/tie/loss counts (%) for our method versus each competing method during the whole active learning process, based on paired t-tests at the 95% confidence level.

Data set	Vs Random	Vs Fbatch	Vs Dbatch	Vs Tbatch	Vs Mbatch
banana	13/ 87/0	39/61/0	80/20/0	69/31/0	17/83/0
diabetis	28/72/0	36/64/0	56/44/0	56/44/0	14/86/0
heart	11/89/0	54/46/0	11/89/0	11/89/0	17/79/4
twonorm	44/56/0	21/76/3	81/17/2	32/66/2	51/49/0
waveform	70/30/0	0/97/3	59/41/0	63/37/0	37/60/2
ringnorm	22/78/0	3/97/0	75/25/0	70/30/0	20/80/0
thyroid	4/96/0	39/61/0	9/91/0	78/22/0	17/83/0
chess	18/82/0	22/78/0	45/55/0	83/17/0	—
ionosphere	18/76/6	24/76/0	9/88/3	59/41/0	—
splice	77/23/0	84/16/0	92/8/0	84/16/0	—
vote	36/62/2	29/71/0	76/22/2	55/40/5	—
image	38/62/0	5/90/5	98/2/0	81/18/1	—
crx	3/97/0	3/97/0	3/97/0	0/100/0	—
australian	0/100/0	0/100/0	2/98/0	0/100/0	—
monk1	28/72/0	34/66/0	28/72/0	28/72/0	—

ing, and a proper balance of these two sources of information will boost the active learning performance.

4.2 Sensitivity

Our algorithm has a tunable parameter β. It balances the trade-off between the effect of discriminative information and representative information in our optimization objective. In this experiment, we run our algorithm with parameter values from a candidate set $\{1, 2, 10, 100, 1000\}$, and show the active learning performance. We report results on two UCI benchmark data sets: breast cancer and german [16]. The experiment settings are the same as previous ones.

We present the results in Figure 2. From these results, we observe that the performance on german is not sensitive to the trade-off parameter. However, the performance on breast cancer is more sensitive to this parameter. The reason may be that the breast cancer data set may have more regular data structure. Therefore, using more representative information helps to boost the active learning performance. In german, the samples may have more complex distribution, which is more difficult to capture. As a result, it does not help much to focus on the representative information. Though these two experiments show different sensitivity behaviors of the parameter, we observe that it does not hurt to pay more attention to the representative information for both data sets. We conclude from these experiments that a relatively larger value of β is recommended, when there are scarce initial labeled samples. In this situation, we need to pay more attention to the data distribution.

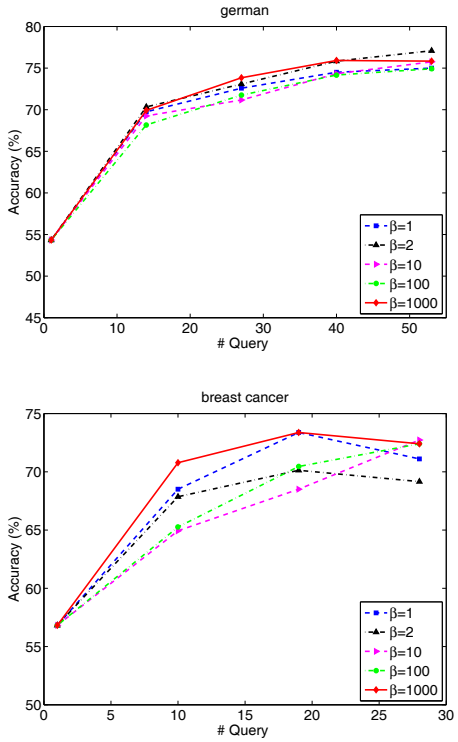

Figure 2: Performance comparison using different trade-off parameters on breast cancer and german data sets for our BMDR algorithm. Each curve represents the average result of 10 runs.

4.3 Discussion

In the active learning literature, both representative information and discriminative information are important for the efficient query. However, they are usually contradictory in the active learning process. Several existing investigations [14, 26, 35, 20] show that the representative information is more useful when there is no or very few labeled data; and the discriminative information is more efficient to boost the learning accuracy when there are certain amounts of labeled data, which can train a classifier with good discriminative capability. Using either information alone may not obtain the best performance during the entire active learning process. In the ideal case, the most efficient active learning method should pay more attention to the representative samples when there are very few labeled samples, and focus

on finding the most discriminative sample to label when the representativeness of the queries decays rapidly.

In this paper, our method accomplishes this goal by properly using those two kinds of information. In the beginning phase, there is no or very few labeled samples, and the empirical risk for the labeled samples is negligible in the optimization objective. In such situation, our method is very similar to the pure representative active learning methods [11, 35]. When the number of labeled samples increases, the discriminative information plays a more and more important role during the queries. When there are sufficient labeled samples, the query of new samples has less effect on the labeled data distribution. The discriminative information begins to play the dominant role. The method becomes more similar to the most discriminative active learning method [20]. With this mechanism, our method naturally balances the effect of the two kinds of information, and makes its queries more efficient.

5. CONCLUSION

In this paper, we generalize the empirical risk minimization principle to the active learning setting and propose a novel active learning method. In our method, we query the samples which are expected to rapidly reduce the empirical risk, and preserve the original source distribution at the same time. This enables our method to achieve consistent good performance during the whole active learning process. We also propose a practical batch mode active learning algorithm which is solved by alternating optimization. The superior performance of our method is verified by our extensive evaluations using benchmark data sets, compared with the state-of-the-art batch mode active learning methods. We observe from our experiments that it is beneficial to update the trade-off parameter which balances the discriminative and representative information during the query process. We plan to develop an adaptive mechanism to tune this parameter automatically, similar to [32]. This could make our active learning framework more practical. In addition, we plan to extend our method to the semi-supervised learning and multi-class learning settings.

6. ACKNOWLEDGMENTS

This research is sponsored in part by NSF CCF-1025177, NIH LM010730, and ONR N00014-11-1-0108.

7. REFERENCES

[1] N. Abe, B. Zadrozny, and J. Langford. Outlier detection by active learning. In *Proceedings of the 12th ACM SIGKDD International Conference in Knowledge Discovery and Data Mining (KDD)*, pages 504–509, 2006.

[2] P. L. Bartlett and M. S. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

[3] A. Beygelzimer, S. Dasgupta, and J. Langford. Importance weighted active learning. In *Proceedings of the 26th International Conference on Machine Learning (ICML)*, pages 49–56, 2009.

[4] J. C. Bezdek and R. J. Hathaway. Convergence of alternating optimization. *Neural, Parallel & Scientific Computations*, 11(4):351–368, 2003.

[5] K. M. Borgwardt, A. Gretton, M. J. Rasch, H.-P. Kriegel, B. Schölkopf, and A. J. Smola. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22(14):49–57, 2006.

[6] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3:1–122, 2011.

[7] C. J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, 1998.

[8] C. Campbell, N. Cristianini, and A. J. Smola. Query learning with large margin classifiers. In *Proceedings of the Seventeenth International Conference on Machine Learning (ICML)*, pages 111–118, 2000.

[9] C.-C. Chang and C.-J. Lin. Libsvm: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27:1–27:27, 2011.

[10] O. Chapelle, B. Schölkopf, and A. Zien, editors. *Semi-Supervised Learning*. MIT Press, Cambridge, MA, 2006.

[11] R. Chattopadhyay, Z. Wang, W. Fan, I. Davidson, S. Panchanathan, and J. Ye. Batch mode active sampling based on marginal probability distribution matching. In *Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 741–749, 2012.

[12] D. A. Cohn, Z. Ghahramani, and M. I. Jordan. Active learning with statistical models. *Journal of Artificial Intelligence Research*, 4(1):129–145, 1996.

[13] F. d'Alché Buc, Y. Grandvalet, and C. Ambroise. Semi-supervised marginboost. In *Advances in Neural Information Processing Systems (NIPS) 14*, 2002.

[14] S. Dasgupta. Two faces of active learning. *Theoretical Computer Science*, 412(19):1767–1781, 2011.

[15] R. M. Dudley. *Real analysis and probability*. Cambridge University Press, 2002.

[16] A. Frank and A. Asuncion. UCI machine learning repository, 2010.

[17] Y. Freund, H. S. Seung, E. Shamir, and N. Tishby. Selective sampling using the query by committee algorithm. *Machine Learning*, 28(2-3):133–168, 1997.

[18] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.

[19] Y. Guo. Active instance sampling via matrix partition. In *Advances in Neural Information Processing Systems (NIPS) 23*, pages 802–810, 2010.

[20] Y. Guo and D. Schuurmans. Discriminative batch mode active learning. In *Advances in Neural Information Processing Systems (NIPS) 20*, pages 593–600, 2008.

[21] S. C. Hoi, R. Jin, J. Zhu, and M. R. Lyu. Batch mode active learning and its application to medical image classification. In *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, pages 417–424, Pittsburgh, PA, USA, 2006.

[22] S.-J. Huang, R. Jin, and Z.-H. Zhou. Active learning by querying informative and representative examples. In *Advances in Neural Information Processing Systems (NIPS) 23*, pages 892–900, 2010.

[23] Y. Koren. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD)*, pages 426–434, 2008.

[24] H. T. Nguyen and A. Smeulders. Active learning using pre-clustering. In *Proceedings of the 21st International Conference on Machine Learning (ICML)*, pages 79–86, New York, NY, USA, 2004. ACM.

[25] N. Roy and A. McCallum. Toward optimal active learning through sampling estimation of error reduction. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML)*, pages 441–448, 2001.

[26] B. Settles. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009.

[27] H. S. Seung, M. Opper, and H. Sompolinsky. Query by committee. In *Proceedings of the Fifth Annual Conference on Computational Learning Theory (COLT)*, pages 287–294, 1992.

[28] B. K. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf, and G. R. Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11:1517–1561, 2010.

[29] M. Sugiyama. Active learning in approximately linear regression based on conditional expectation of generalization error. *Journal of Machine Learning Research*, 7:141–166, 2006.

[30] S. Tong and D. Koller. Support vector machine active learning with applications to text classification. *Journal of Machine Learning Research*, 2:45–66, 2002.

[31] V. Vapnik. *Statistical learning theory*. Wiley, 1998.

[32] Z. Wang, S. Yan, and C. Zhang. Active learning with adaptive regularization. *Pattern Recognition*, 44(10-11):2375–2383, 2011.

[33] M. K. Warmuth, G. Rätsch, M. Mathieson, J. Liao, and C. Lemmen. Active learning in the drug discovery process. In *Advances in Neural Information Processing Systems (NIPS) 14*, pages 1449–1456, 2001.

[34] Z. Xu, K. Yu, V. Tresp, X. Xu, and J. Wang. Representative sampling for text classification using support vector machines. In *Proceedings of European Conference on Information Retrieval (ECIR)*, pages 393–407, 2003.

[35] K. Yu, J. Bi, and V. Tresp. Active learning via transductive experimental design. In *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, pages 1081–1088, 2006.

[36] X. Zhu, J. Lafferty, and Z. Ghahramani. Combining active learning and semi-supervised learning using gaussian fields and harmonic functions. In *ICML 2003 workshop on The Continuum from Labeled to Unlabeled Data in Machine Learning and Data Mining*, 2003.

APPENDIX

A. PROOF OF THEOREM 2.1

PROOF. Following [18], we know that the relationship between the true MMD and the empirical MMD is

$$Pr\left(\left|\mathrm{MMD}[\mathcal{C}, p(\mathbf{x}), q(\mathbf{x})] - \mathrm{MMD}_\phi(S, Q)\right| \geq \epsilon + 2(\sqrt{\tfrac{M}{n}} + \sqrt{\tfrac{M}{q}})\right)$$
$$\leq 2e^{\frac{-\epsilon^2 nq}{2M(n+q)}}.$$

with the empirical MMD term given by

$$\mathrm{MMD}_\phi(S, Q) = \left\| \frac{1}{n}\sum_{\mathbf{x}_i \in S} \phi(\mathbf{x}_i) - \frac{1}{q}\sum_{\mathbf{x}_i \in Q} \phi(\mathbf{x}_i) \right\|_{\mathcal{F}}.$$

In the active leaning scenario, $Q \subseteq S$ and $q \leq n$. We have

$$e^{\frac{-\epsilon^2 nq}{2M(n+q)}} \leq e^{\frac{-\epsilon^2 nq}{2M(n+n)}}$$

and

$$\sqrt{\frac{M}{n}} + \sqrt{\frac{M}{q}} \geq 2\sqrt{\frac{M}{n}}.$$

Then

$$Pr\Big(\mathrm{MMD}[\mathcal{C}, p(\mathbf{x}), q(\mathbf{x})] \geq \mathrm{MMD}_\phi(S, Q) + \epsilon + 4\sqrt{\tfrac{M}{n}}\Big)$$
$$\leq 2e^{\frac{-\epsilon^2 q}{4M}}.$$

Let $2e^{\frac{-\epsilon^2 q}{4M}} = \delta/2$. We obtain $\epsilon = \sqrt{\frac{4M\ln(4/\delta)}{q}}$.

From the analysis in Section 2, we know by the classic ERM principle that

$$\mathrm{E}_D(l(f(\mathbf{x}), y)) \leq \hat{\mathrm{E}}_Q(l(f(\mathbf{x}), y)) + \mathrm{MMD}[\mathcal{C}, p(\mathbf{x}), q(\mathbf{x})] + \left[2R_q(\mathcal{L}) + \sqrt{\tfrac{\ln(2/\delta)}{q}}\right],$$

holds with probability at least $1 - \delta/2$.

Combining all the results above, we show that with probability at least $1 - \delta$, the following holds:

$$\mathrm{E}_D(l(f(\mathbf{x}), y)) \leq \hat{\mathrm{E}}_Q(l(f(\mathbf{x}), y)) + \mathrm{MMD}_\phi(S, Q) + C(\mathcal{L}, q, \delta).$$

The function complexity term is

$$C(\mathcal{L}, q, \delta) = 2R_q(\mathcal{L}) + \sqrt{\frac{\ln(2/\delta)}{q}} + 4\sqrt{\frac{M}{n}} + \sqrt{\frac{4M\ln(4/\delta)}{q}}.$$

It can be rewritten as:

$$C(\mathcal{L}, q, \delta) = 2R_q(\mathcal{L}) + c\sqrt{\frac{M\ln(1/\delta)}{q}},$$

where c is a constant. □