Learning by Actively Querying Strong Modal Features*

Yang Yang and **De-Chuan Zhan** and **Yuan Jiang**
National Key Laboratory for Novel Software Technology, Nanjing University,
Collaborative Innovation Center of Novel Software Technology and Industrialization
Nanjing, 210023, China
{yangy, zhandc, jiangy}@lamda.nju.edu.cn

Abstract

Complex objects are usually with multiple modal features. In multi-modal learning, modalities closely related to the target tasks are known as strong modalities. While collecting strong modalities of all instances is often expensive, and current multi-modal learning techniques hardly take the strong modal feature extraction expenses into consideration. On the other hand, active learning is proposed to reduce the labeling expenses by querying the ground truths for specific selected instances. In this paper, we propose a training strategy, ACQUEST (ACtive QUErying STrong modalities), which exploits strong modal information by actively querying the strong modal feature values of "selected" instances rather than their corresponding ground truths. In ACQUEST, only the informative instances are selected for strong modal feature acquisition. An inverse prediction technique is also proposed to make the ACQUEST a unified optimization form. Experiments on image datasets show that ACQUEST achieves better classification performance than conventional active learning and multi-modal learning methods with less feature acquisition costs and labeling expenses.

1 Introduction

With the fast development of data collection techniques, complicated objects can be described by features from different data channels and are naturally with multi-modal feature presentations, e.g., modern mobile phones with different type of sensors can collect sensor signals from multiple channels. Recently, multi-modal learning techniques have been developed and paid more attentions to utilize the information from different modalities. Kiros *et al.* [2014] applied deep network to learn features over multiple modal data; Zhou *et al.* [2015] used the multi-modal time-series signals in mental health system; Nguyen *et al.* [2013] proposed the M3LDA to annotate image regions together with text tags, and provided a promising way to understand the relation between input patterns and output semantics. Meanwhile, different modalities are of various importance under specific circumstances, e.g., in medical tests for diagnosing the same disease, it is known that CT tests are with higher confidence, while on the contrary, X-rays are with lower confidence. We denote the modalities which can help the tasks more as strong modalities and on the contrary are the weak modalities. It is notable that strong modalities can lead to better performance, nevertheless, are more expensive for collection or extraction. Although existing modern multi-modal learning techniques can achieve better performance by incorporating with more strong modal features [Yang *et al.*, 2015], they do not take the feature extraction expenses into consideration. Thus, how to reduce the informative strong modal feature extraction costs is an urgent problem.

Active learning aims at learning concepts by querying the labels of unlabeled data for better classification performance as well as reducing the costs of labeling. Settles [2010] designed a learning algorithm that connects active learning with multi-armed bandit; Zhong *et al.* [2015] added the "unsure" option for the crowd annotators in active learning; Huang *et al.* [2015] proposed a novel MLAL framework to query the relevance ordering of label pairs. Yet querying the ground truths is more expensive than collecting features from strong modalities. As a matter of fact, the ground truths can be regarded as a special type of "strong" modal features in certain extents, yet they are gathered by oracle labeling and are supposed with more expenses than strong modal feature collection or extraction by sensors.

In this work, we propose the ACQUEST (ACtive QUErying STrong modalities) strategy, which makes full use of strong modalities by actively querying strong modal features while reducing the feature acquisition expenses simultaneously following the style of active querying. Different from active learning, ACQUEST exploits multiple modalities by querying the features of "selected" instances from strong modalities rather than querying corresponding ground truths. At the same time, in ACQUEST, only the informative instances are selected to query strong modal features. It is notable that ACQUEST queries feature values of strong modalities without any interventions from oracle, and this makes ACQUEST can be trained with the least overall costs. An *inverse prediction* technique is also proposed and embedded for making the ACQUEST unified in one optimization formulation. The ef-

*This research was supported by NSFC (61273301) and CCF-Tencent Open Research Fund (RAGR20150117).

fectiveness of the ACQUEST is validated by extensive experiments. Section 2 is related work, our approach is presented in Section 3. Section 4 reports our experiments. Finally, Section 5 gives the conclusion.

2 Related Work

Multi-Modal learning has attracted many attentions [Wang *et al.*, 2015], [Zhou *et al.*, 2005], [Zhang and Li, 2014] and [Zhang *et al.*, 2014]. Recently, some approaches distinguish the multiple modalities into strong modalities and weak modalities according to their importance for the concrete application requirements [Ion *et al.*, 2006]. By incorporating strong modal information, better weak modal feature extraction can be performed and can consequently achieve better performance [Yang *et al.*, 2015]. Nevertheless, these methods have not taken the expenses of strong modal feature extraction into consideration.

Active learning exploits unlabeled data by querying the labels of a subset of unlabeled instances for better classification performance, and can reduce the oracle labeling costs. The most popular active learning approaches usually choose the informative or representative instances for the ground truth querying [Sachan *et al.*, 2015], [Fang *et al.*, 2014], [Maria-Florina *et al.*, 2007] and [Sanjoy and Daniel, 2008]. Huang *et al.* [2014] presented a novel active learning approach which considers both the informative and representative criteria simultaneously and achieved better classification accuracy while reducing the labeling costs.

Nevertheless, in multi-modal learning scenarios, two different facts induce further considerations on multi-modal feature collection and extraction:

- Strong modal features which help a concerned task more are with higher collection expenses than ordinary weak modal features, e.g., acquisition of finger prints needs specialized equipments;
- The expenses of gathering strong modal features are accomplished with sensors automatically and surely with less manual interactions, therefore are less expensive than directly querying the oracles.

To the best of our knowledge, previous multi-modal methods, which improved the classification performance using the auxiliary strong modalities, did not consider the collection expenses of strong modal features, while active learning methods query the labels of unlabeled instances to reduce the costs of labeling, yet cannot directly applied in the multi-modal learning scenarios. In this work, we focus on actively acquiring the most informative strong modal features from a portion of instances instead of the ground truths to ulteriorly reduce the data acquisition costs. The proposed approach ACQUEST (ACtive QUErying STrong modalities) selects the most informative and representative instances for strong modal feature values acquisition, and there is no need to query oracles for labels. Consequently, ACQUEST results in further reduction of the overall expenses on the collection of features or labels.

3 Proposed method

Suppose we have N examples, denoted by $\mathcal{D} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_{n_l}, y_{n_l}), \mathbf{x}_{n_l+1}, \cdots, \mathbf{x}_N\}$, the training dataset that consists of n_l labeled data D_l and $n_u = N - n_l$ unlabeled data D_u, where each instance $\mathbf{x}_i = [x_{i_1}, x_{i_2}, \cdots, x_{i_d}] \in \mathbb{R}^d$, and $y_i \in \{-1, +1\}$ is the class label of $\mathbf{x}_i$. Meanwhile, in multi-modal learning, instance space can be denoted as, at least two parts without overlap, $v = \{v_1, v_2\}$, where $v_1 \in \mathbb{R}^{d_1}$ is raw features from weak modality and $v_2 \in \mathbb{R}^{d_2}$ is the strong modal raw features, $d = d_1 + d_2$. In this paper, without any loss of generalities, each instance $\mathbf{x}_i$ is denoted as $(\mathbf{x}_{i,v_1}, \mathbf{x}_{i,v_2})$.

3.1 Active Querying Strong Modalities (ACQUEST)

Researchers claim strong modal features, which are with more discriminative abilities, meanwhile are with higher costs [Yang *et al.*, 2015]. Yet it is a matter of fact that the ground truths are more expensive since collecting labels requires human efforts. As a consequence, in case of the connections between the strong modal features and the ground truth concepts are "idea" exploited, we can actively query the strong modal features of some informative and representative instances rather than true labels, to reduce the labeling expenses. This will result in less costs for feature value acquisition and is the basic idea of the proposed ACQUEST (ACtive QUErying STrong modalities). In this section, we focus on describing the novel approach in detail.

We start our discussion from a regularized classifier which can be generally formed as:

$$f^* = \operatorname*{argmin}_{f \in \mathcal{H}} \frac{\lambda_1}{2} \|f\|^2_{\mathcal{H}} + \sum_{i=1}^{n_l} \ell(y_i, f(\mathbf{x}_{i,v_1})),$$

where f^* can be a linear or kernelized classifier trained from labeled examples, $\mathcal{H}$ is the Reproducing Kernel Hilbert Space, $\ell(z, \hat{z})$ can be any convex loss function. Given the classifier f^*, the labels of unlabeled instances which are close to the decision boundary should be selected for ground truth querying in active learning, i.e., the selected instance $\mathbf{x}_{s,v_1}$ should lead to a small value for the object function regardless of its label y_s.

In active learning scenarios, in order to select queries that are more representative [Huang *et al.*, 2014], the evaluation function can be extended to include all the unlabeled data. If we know the class assignments $\mathbf{y}_u \in \{\pm 1\}^{n_u - 1}$ for all unselected unlabeled instances in $\mathcal{D}_u$, the criterion of instances selection can be approximated by:

$$s^* = \operatorname*{argmin}_{n_l < s \leq N} L(\mathcal{D}_l, \mathcal{D}_u, \mathbf{y}_u, \mathbf{x}_{s,v_1}),$$

where the evaluation function can be represented as:

$$L(\mathcal{D}_l, \mathcal{D}_u, \mathbf{y}_u, \mathbf{x}_{s,v_1}) = \max_{y_s} \min_{f_{v_1} \in \mathcal{H}} \frac{\lambda_1}{2} \|f_{v_1}\|^2_{\mathcal{H}} + \sum_{i=1}^{N} \ell(y_i, f_{v_1}(\mathbf{x}_{i,v_1})),$$

To select more informative example $\mathbf{x}_{s,v_1}$, we expect that all unselected unlabeled instances from $\mathcal{D}_u$ should result in a small value of $L(\mathcal{D}_l, \mathcal{D}_u, \mathbf{y}_u, \mathbf{x}_{s,v_1})$ in contrary to y_s maximizing the value of $L(\mathcal{D}_l, \mathcal{D}_u, \mathbf{y}_u, \mathbf{x}_{s,v_1})$. We therefore approximate the solution for $\mathbf{y}_u$ by minimizing $L(\mathcal{D}_l, \mathcal{D}_u, \mathbf{y}_u, \mathbf{x}_{s,v_1})$, which leads to the following surrogate for $L(\mathcal{D}_l, \mathcal{D}_u, \mathbf{y}_u, \mathbf{x}_{s,v_1})$ in query selection:

$$\hat{L}(\mathcal{D}_l, \mathcal{D}_u, \mathbf{x}_{s,v_1}) = \min_{\mathbf{y}_u} \max_{y_s} \min_{f_{v_1} \in \mathcal{H}} \frac{\lambda_1}{2} \|f_{v_1}\|^2_{\mathcal{H}} + \sum_{i=1}^{N} \ell(y_i, f_{v_1}(\mathbf{x}_{i,v_1})),$$

where $\mathbf{y}_u \in \{\pm 1\}^{n_u - 1}$, $y_s \in \{\pm 1\}$ and eventually we can reduce the labeling expenses using active queries on labels.

While in multi-modal learning scenarios, gathering ground truths should be more complicated, i.e., labeling multi-modal instances requires oracles inspecting more feature values from different modalities. However, by noticing that the expenses of collecting strong modal feature values (usually automatically collected by sensors or computers) are greatly less than that of labeling, we then can turn to query strong modal feature values instead of ground truths. ACQUEST approach is used to achieve this goal. Through explicitly figuring out the connections between strong modality and ground truth concepts, ACQUEST can integrate an inverse prediction technique to unify actively querying with multi-model learning in one optimization framework.

To simplify the discussion, we first model the connections between strong modal features $\mathbf{x}_{i,v_2}$ and weak modalities $\mathbf{x}_{i,v_1}$ with a linear model. It is assumed there can be an appropriate feature subspace on strong modality homologous with weak modal feature space: $\hat{\mathbf{x}}_{i,v_2} = g(\mathbf{x}_{i,v_1}) = \mathbf{x}_{i,v_1} W$, where the weak raw feature projection matrix $W \in \mathbb{R}^{d_1 \times d_2}$. Note that this kind of connection can be easily extended to kernelized version further. The corresponding loss function describing the difference between strong modalities and transformed weak features is defined as: $\ell'(\mathbf{x}_{i,v_2}, g(\mathbf{x}_{i,v_1}))$. Here, $\ell'(z, \hat{z})$ can be any convex loss functions. Similar to the definition of $\hat{L}(\mathcal{D}_l, \mathcal{D}_u, \mathbf{x}_{s,v_1})$, we have

$$\begin{aligned}\hat{G}(D_l, D_u, \mathbf{x}_{s,v_1}) = & \min_{\mathbf{x}_{u,v_2} \in X_{u,v_2}} \max_{\mathbf{x}_{s,v_2}} \min_{g_{v_1} \in \mathcal{H}} \frac{\lambda_2}{2} \|g_{v_1}\|_{\mathcal{H}}^2 \\ & + \sum_{i=1}^{N} \ell'(\mathbf{x}_{i,v_2}, g(\mathbf{x}_{i,v_1})),\end{aligned}$$

where $X_{u,v_2} = [\mathbf{x}_{n_l+1,v_2}, \mathbf{x}_{n_l+2,v_2}, \ldots, \mathbf{x}_{N,v_2}] \in \mathbb{R}^{(n_u-1) \times d_2}$ is the unlabeled strong modal features except for $\mathbf{x}_{s,v_2}$ which is the strong modal feature values corresponding to the instance queried.

For simplicity, in our implementation, we set $\ell(y, \hat{y}) = (y - \hat{y})^2/2$ and model the connection between strong modal features and ground truth concepts linearly, i.e., we assume $\mathbf{y}_l = X_{l,v_2}\mathbf{w}$, where the linear coefficients $\mathbf{w} \in \mathbb{R}^{d_2}$. X_{l,v_2} is the strong modal features of labeled examples. In this way, the connections between strong modal features and labels are actually in the form of least square minimization: $\operatorname{argmin} \|X_{l,v_2}\mathbf{w} - \mathbf{y}_l\|_2^2$, and of course has a closed form solution: the strong modal feature values can be represented as $X_{l,v_2} = \mathbf{y}_l \mathbf{w}^\dagger$, where $\mathbf{w}^\dagger$ is the pseudo inverse of $\mathbf{w}$. By replacing the strong modal instance feature values X_{v_2} with $\mathbf{y}\mathbf{w}^\dagger$, we can reform the equation above as:

$$\begin{aligned}\hat{G}(\mathcal{D}_l, \mathcal{D}_u, \mathbf{x}_{s,v_1}) = & \min_{\mathbf{y}_u} \max_{y_s} \min_{g_{v_1} \in \mathcal{H}} \frac{\lambda_2}{2} \|g_{v_1}\|_{\mathcal{H}}^2 \\ & + \sum_{i=1}^{N} (y_i \mathbf{w}^\dagger - g(\mathbf{x}_{i,v_1}))^2.\end{aligned} \tag{1}$$

It is notable that the closed form solution of $X_{l,v_2} = \mathbf{y}_l \mathbf{w}^\dagger$ provides an "*inverse prediction*" of strong modal features X_{v_2} with (pseudo) labels $\mathbf{y}$ and bridges the strong modalities and ground truths, which will further unify the ACQUEST in a holistic framework. In particular, to choose the most informative instances to query the strong modal feature values, we should consider both the $\hat{L}(\mathcal{D}_l, \mathcal{D}_u, \mathbf{x}_{s,v_1})$ and $\hat{G}(\mathcal{D}_l, \mathcal{D}_u, \mathbf{x}_{s,v_1})$ simultaneously, and eventually the query of ACQUEST is made according to the following criterion:

$$\begin{aligned}\hat{A}(\mathcal{D}_l, \mathcal{D}_u, \mathbf{x}_{s,v_1}) = & \hat{L}(\mathcal{D}_l, \mathcal{D}_u, \mathbf{x}_{s,v_1}) + \hat{G}(\mathcal{D}_l, \mathcal{D}_u, \mathbf{x}_{s,v_1}) \\ = & \min_{\mathbf{y}_u} \max_{y_s} \min_{f_{v_1}, g_{v_1}} \frac{\lambda_1}{2} \|f_{v_1}\|_{\mathcal{H}}^2 + \frac{\lambda_2}{2} \|g_{v_1}\|_{\mathcal{H}}^2 \\ & + \frac{1}{2} \|\mathbf{y} - F_{v_1}\|_F^2 + \frac{1}{2} \|\mathbf{y}\mathbf{w}^\dagger - G_{v_1}\|_F^2,\end{aligned}$$

and the instance need to be queried can be obtained by:

$$s^* = \operatorname*{argmin}_{n_l < s \leq N} \hat{A}(\mathcal{D}_l, \mathcal{D}_u, \mathbf{x}_{s,v_1}), \tag{2}$$

where $F_{v_1} = \{f_{v_1}(x_{1,v_1}), f_{v_1}(x_{2,v_1}), \ldots, f_{v_1}(x_{N,v_1})\} \in \mathbb{R}^N$ is the predictors for features of weak modality, and $G_{v_1} = \{g_{v_1}(x_{1,v_1}), g_{v_1}(x_{2,v_1}), \ldots, g_{v_1}(x_{N,v_1})\} \in \mathbb{R}^{N \times d_2}$ is the extracted features from weak modality. $\mathbf{y} = [y_1, y_2, \ldots, y_N] \in \mathbb{R}^N$, where for very limited number of the labeled data, $y_i = 1$ if $\mathbf{x}_i$ belongs to positive class, and $y_i = -1$ is negative. For unlabeled instances, $\mathbf{y}_u$ equals to 1 or -1. It is notable that in Eq. 2, the strong modal features are represented by $\mathbf{y}_u$ and $\mathbf{w}^\dagger$ with the "*inverse predictions*" treatments which simplifies the enumeration of feature values for strong modalities.

$\|f_{v_1}\|_{\mathcal{H}}^2$ and $\|g_{v_1}\|_{\mathcal{H}}^2$ are the structure risk of predictor f_{v_1} and feature extractor g_{v_1} in each function space respectively. In order to predict with the lower-cost weak modal information only in test phase, we simply assume a linear predictor defined in the extracted feature space of weak modalities, i.e., $F_{v_1} = X_{v_1} W \mathbf{v}$, where $\mathbf{v} \in \mathbb{R}^{d_2}$. As a consequence, the definition of $\hat{A}(\mathcal{D}_l, \mathcal{D}_u, \mathbf{x}_{s,v_1})$ can be reformed as:

$$\begin{aligned}\min_{\mathbf{y}_u} \max_{y_s} \min_{W, \mathbf{v}} & \frac{\lambda_1}{2} \|f_{v_1}\|_{\mathcal{H}}^2 + \frac{1}{2} \|\mathbf{y} - X_{v_1} W \mathbf{v}\|_F^2 + \\ & \frac{\lambda_2}{2} \|g_{v_1}\|_{\mathcal{H}}^2 + \frac{1}{2} \|\mathbf{y}\mathbf{w}^\dagger - X_{v_1} W\|_F^2,\end{aligned} \tag{3}$$

where the weak modal instances X_{v_1} can be projected into the strong modal feature space and represented as $X_{v_1} W$, therefore the 4th term reduces the differences between strong modal features and projected weak features. $W\mathbf{v}$ is the predictor for X_{v_1}, consequently $\|\mathbf{y} - X_{v_1} W \mathbf{v}\|_F^2$ can act as linear classifier on weak modality. Note that different from [Yang *et al.*, 2015], ACQUEST can reduce the strong modal feature costs on both test and training phase, while the former needs all strong modal features during training.

4 Solution to Training ACQUEST Model

In this section, we further derive the training approach for selecting the strong modal features of the instance which should be queried. When least square loss is utilized, it is straightforward that

$$\min_{f_{v_1} \in \mathcal{H}} \frac{\lambda_1}{2} \|f_{v_1}\|_{\mathcal{H}}^2 + \frac{1}{2} \|\mathbf{y} - F_{v_1}\|_F^2 = \frac{1}{2} \mathbf{y}^\top L^{v_1, l} \mathbf{y},$$

where $L^{v_1,l} = (\lambda_1 I + X_{v_1}^\top X_{v_1})^{-1}$. As a consequence, one part of target function $\hat{L}(\mathcal{D}_l, \mathcal{D}_u, \mathbf{x}_{s,v_1})$ can be simplified as:

$$\hat{L}(\mathcal{D}_l, \mathcal{D}_u, \mathbf{x}_{s,v_1}) = \min_{\mathbf{y}_u} \max_{y_s} \mathbf{y}^\top L^{v_1, l} \mathbf{y}.$$

Similarly, for the classification model trained by the strong modal examples, we also can have:

$$\min_{g_{v_1} \in \mathcal{H}} \frac{\lambda_2}{2} \|g_{v_1}\|_{\mathcal{H}}^2 + \frac{1}{2} \|\mathbf{y}\mathbf{w}^\dagger - G_{v_1}\|_F^2 = \frac{1}{2} \operatorname{tr}(\hat{\mathbf{y}}^\top L^{v_1, v_2} \hat{\mathbf{y}}),$$

where $L^{v_1,v_2} = (\lambda_2 I + X_{v_1}^\top X_{v_1})^{-1}$ and $\hat{\mathbf{y}} = \mathbf{y}\mathbf{y}^\top X_{v_2}(X_{v_2}^\top X_{v_2})^{-1}$, and $\hat{G}(\mathcal{D}_l, \mathcal{D}_u, \mathbf{x}_{s,v_1})$ can be simplified as:

$$\hat{G}(\mathcal{D}_l, \mathcal{D}_u, \mathbf{x}_{s,v_1}) = \min_{\mathbf{y}_u} \max_{y_s} \operatorname{tr}(\hat{\mathbf{y}}^\top L^{v_1,v_2} \hat{\mathbf{y}}).$$

According to Eq. 2, we have:

$$\hat{A}(D_l, D_u, x_s) = \min_{\mathbf{y}_u} \max_{y_s} \mathbf{y}^\top L^{v_1,l} \mathbf{y} + \operatorname{tr}(\hat{\mathbf{y}}^\top L^{v_1,v_2} \hat{\mathbf{y}}). \tag{4}$$

Query the most representative and informative instance needs to compute the evaluation function $\hat{A}(D_l, D_u, x_s)$ in Eq. 4 for every unlabeled instance $\mathbf{x}_{s,v_1}$. Note that the calculation can be further reduced or simplified since

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

where L can be $L^{v_1,l}$ or L^{v_1,v_2}. $L_{m,n}$ is the sub-matrix of L with corresponding rows and columns according to subscripts m and n, and the above objective is concave in y_s and convex in $\mathbf{y}_u$, the minimization on $\mathbf{y}_u$ and maximization on y_s can be switched. Moreover, the solution to $\min\limits_{\mathbf{y}_u} \mathbf{y}_u^\top L \mathbf{y}_u$ can be obtained in a closed form:

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

By substitute $\hat{\mathbf{y}}_u$ to $\hat{A}(D_l, D_u, x_s)$, we consequently have

$$\begin{aligned} \hat{A}(D_l, D_u, x_s) = & L_{s,s}^{v_1,l} + \mathbf{y}_l^\top L_{l,l}^{v_1,l} \mathbf{y}_l + L_{s,s}^{v_1,v_2} + \operatorname{tr}(\hat{\mathbf{y}}_l^\top L_{l,l}^{v_1,v_2} \hat{\mathbf{y}}_l) \\ & + \max_{y_s} \Big[- (L_{u,l}^{v_1,l} \mathbf{y}_l + L_{u,s}^{v_1,l} y_s)^\top L_{u,u}^{v_1,l\,-1} (L_{u,l}^{v_1,l} \mathbf{y}_l + L_{u,s}^{v_1,l} y_s) \\ & - \operatorname{tr}[(L_{u,l}^{v_1,v_2} \hat{\mathbf{y}}_l + L_{u,s}^{v_1,v_2} y_s)^\top L_{u,u}^{v_1,v_2\,-1} (L_{u,l}^{v_1,v_2} \hat{\mathbf{y}}_l + L_{u,s}^{v_1,v_2} y_s)] \\ & + 2 y_s L_{s,l}^{v_1,l} \mathbf{y}_l + \operatorname{tr}(2 y_s L_{s,l}^{v_1,v_2} \hat{\mathbf{y}}_l) \Big]. \end{aligned}$$

Thus in ACQUEST training phase, we need to evaluate the criterion in Eq. 5 to select the instance for strong modal feature values querying according to Eq. 2, this procedure should be repeated for unlabeled instances iteratively.

In each iteration, one most valuable unlabeled instance is selected for strong modal feature value acquisition. Once we have obtained the strong modal features of the most valuable instance, we can have the parameters W and $\mathbf{v}$ updated. Note that the 2nd term in Eq. 3 involves the product of weak modal feature extraction matrix W and the weak modal predictor $\mathbf{v}$, we can use the gradient descent techniques for updating the parameters. Here alternative descent algorithm is used in this work, since this will lead to closed form alternating and finally simplify the calculations.

Fix W, Update $\mathbf{v}$

When W is fixed, note that the 3rd and 4th term in Eq. 3 are not related to $\mathbf{v}$, therefore it can be equivalently written as:

$$\operatorname*{argmin}_{\mathbf{v}} \frac{\lambda_1}{2} \mathbf{v}^\top W^\top W \mathbf{v} + \frac{1}{2} \|\mathbf{y} - X_{v_1} W \mathbf{v}\|_F^2 \tag{5}$$

Eq. 5 has a closed-form solution for $\mathbf{v}$:

$$\mathbf{v} = (W^\top (\lambda_1 I + X_{v_1}^\top X_{v_1}) W)^{-1} W^\top X_{v_1}^\top \mathbf{y}.$$

Algorithm 1 The ACQUEST Algorithm

```
Require: $X_{l,v_1}$, $X_{u,v_1}$, $X_{l,v_2}$, $X_{u,v_2}$, $\lambda_1$, $\lambda_2$, $\mathbf{y}$, max-iter;
repeat
    for $i = 1$ to $n_u$ do
        Calculate $\hat{A}(\mathcal{D}_l, \mathcal{D}_u, \mathbf{x}_{s,v_1})$ using Eq. 5
    end for
    Select the $\mathbf{x}_{s^*}$ with the smallest $\hat{A}(\mathcal{D}_l, \mathcal{D}_u, \mathbf{x}_{s,v_1})$
    Calculate the pseudo label $y_{s^*}$ using $y = \text{sign}(\mathbf{x}_{s,v_2} \mathbf{w})$
    $X_{l,v_1} = X_{l,v_1} \cup (\mathbf{x}_{s^*}, y_{s^*})$; $X_{u,v_1} = X_{u,v_1} \setminus \mathbf{x}_{s^*}$
    Training classifier with closed solutions of Eq. 5 and 6
until the query number (number of iterations) exceeds max-iter.
```

Fix $\mathbf{v}$, Update W

When $\mathbf{v}$ is fixed, Eq. 3 can be equivalently written as:

$$\begin{aligned} \operatorname*{argmin}_{W} & \frac{\lambda_1}{2} \mathbf{v}^\top W^\top W \mathbf{v} + \frac{1}{2} \|\mathbf{y} - X_{v_1} W \mathbf{v}\|_F^2 + \\ & \frac{\lambda_2}{2} W^\top W + \frac{1}{2} \|\mathbf{y}\mathbf{w}^\dagger - X_{v_1} W\|_F^2 \end{aligned} \tag{6}$$

Obviously, Eq. 6 also has a closed solution for W.

Prediction with only weak modality required

In the prediction stage, we can predict the test instances which are represented with weak modalities only (i.e., no strong modal features are provided), with the obtained W and $\mathbf{v}$ in training phase in following equation:

$$f_{v_1}(\mathbf{x}_{i,v_1}) = \mathbf{x}_{i,v_1} W \mathbf{v} + b,$$

where $\mathbf{x}_{i,v_1}$ is weak feature values of a test instance. The bias b for predictors is obtained during the training stage, and the instance label should depend on $\text{sign}(f_{v_1}(\mathbf{x}_{i,v_1}))$. The pseudo code of ACQUEST is shown in Algorithm 1.

5 Experiment

Datasets and Configurations

In this section, we introduce the compared methods and datasets before giving the empirical results of ACQUEST. ACQUEST can be adopted for many applications where there are multi-modal features. In this paper, 12 image datasets are used in our empirical investigations. A subset of NUS [Chua *et al.*, 2009] contains 9,109 images of 10 categories, and 6 groups of features extracted. MSRA [Wang *et al.*, 2009] subset contains 10,680 images of 9 categories, and 7 groups of features are extracted. Animal [Christoph *et al.*, 2009] (represented as ANIM in short in the following content) subset contains 30475 images of 50 animals classes, and 6 pre-extracted feature representations are extracted for each image. All the feature sets can be separated into strong modal features and weak modality. More specifically, for NUS the color histogram features are selected as weak modality while the rest are strong modal features, and in NUS, 4 subsets are constructed for balanced binary classification, i.e., lake vs. railroad, surf vs. map, map vs. boats and reflection vs. boats (denoted as NUS_1, NUS_2, NUS_3 and NUS_4 respectively). Similarly, for MSRA, HSV color histogram is the weak modal features and the rest are strong modal features, and 4 balanced subsets are selected for binary classification, denoted as MSRA_1, MSRA_2, MSRA_3 and MSRA_4 respectively. For

Table 1: The accuracies (avg.±std.) compared to active learning and multi-modal learning approaches. The best classification performance is bolded.

	ACQUEST	MARGIN	RANDOM	QUIRE	IDE	DUAL	ARM
NUS$_1$	**.786±.011**	.747±.004	.756±.002	.756±.002	.756±.003	.732±.011	.726±.091
NUS$_2$	**.748±.052**	.709±.033	.715±.037	.678±.023	.678±.037	.717±.036	.652±.014
NUS$_3$	**.832±.033**	.832±.037	.814±.030	.802±.021	.802±.026	.806±.026	.783±.060
NUS$_4$	.669±.042	.599±.008	.627±.026	.621±.024	.621±.028	.605±.024	**.719±.076**
MSRA$_1$	.692±.018	.667±.022	.664±.017	.584±.042	.584±.037	.571±.044	**.752±.077**
MSRA$_2$	.750±.012	**.777±.016**	.749±.008	.739±.003	.739±.009	.718±.010	.625±.074
MSRA$_3$	**.771±.031**	.744±.018	.724±.020	.678±.036	.678±.005	.739±.012	.682±.141
MSRA$_4$	**.772±.009**	.749±.003	.712±.007	.756±.005	.756±.003	.657±.026	.601±.016
ANIM$_1$	**.643±.005**	.626±.003	.613±.007	.620±.004	.640±.011	.613±.006	.574±.015
ANIM$_2$	**.804±.004**	.803±.009	.777±.002	.788±.002	.788±.005	.782±.003	.599±.023
ANIM$_3$	**.721±.032**	.720±.018	.690±.007	.707±.016	.707±.014	.654±.006	.585±.033
ANIM$_4$	.786±.008	**.791±.011**	.776±.008	.766±.005	.766±.010	.752±.003	.630±.000

ANIM, PyramidHOG (PHOG) features are weak modal features and the rest are strong modal features. Besides, 4 subsets are also constructed from ANIM, i.e., horse vs. cow, rhinoceros vs. otter, rhinoceros vs. collie and rhinoceros vs. raccoon (denoted as ANIM$_1$, ANIM$_2$, ANIM$_3$ and ANIM$_4$). The judgement of weak/strong modalities are made according to the feature extraction time costs of each group modalities.

For all datasets, 66% instances are randomly picked up for training, and the remains are used as test set. The labeled ratio is set to 10% for training set. All experiments are repeated for 30 times. During the training phase, at most 30 unlabeled instances are automatically selected for strong modal feature value querying in each iteration. More specifically, in ACQUEST an unlabeled data instance is first selected for strong modal feature values querying according to the criterion discussed in Eq. 5, followed by the calculations of the corresponding pseudo labels, and then the classifier is retrained using this corresponding instance with both strong modal features and its pseudo label. The avg. and std. of predictions are recorded for evaluation. In all experiments, the parameters λ_1 and λ_2 in the training phase are tuned in $\{10^{-1}, 1, 10\}$. Empirically ACQUEST converges when the difference of the objective value of Eq. 3 is less than 10^{-5}.

Since ACQUEST actively queries the strong modal feature values, we compare it with 5 active learning algorithms and a recently proposed semi-supervised multi-modal learning approach [Yang *et al.*, 2015] which also takes strong/weak modalities in consideration. For 5 compared active learning algorithms, both strong and weak modal features are provided and these active learning models are trained with ground truths while queries are made. The 6 compared methods are:
RANDOM: active learning with randomly selected queries;
MARGIN: margin-based active learning, linear classifier as the base learner [Tong and Koller, 2002];
IDE: active learning that selects informative and diverse examples [Jin *et al.*, 2008];
DUAL: a dual strategy for active learning that exploits both informativeness and representativeness for selection [Pinar *et al.*, 2007]. Parameter of DUAL is set as $k = 1$;
QUIRE: active learning algorithm which queries unlabeled instances that are both informative and representative;
ARM: multi-modal learning approach which improves multi-modal learning performance via extracting the most discriminative weak modal feature subspace with the help of strong modal information.

Classification Accuracy Comparisons

Table 1 records the accuracies (avg.± std.) of the ACQUEST and compared methods, the number of queries is 30 for active learning methods and ACQUEST. For each dataset, the best result is highlighted in bold, ACQUEST is tested with f_{v_1} on weak modality only.

From Table 1, it clearly reveals that on 8 real world datasets, the average accuracies of ACQUEST are the best, and for the rest datasets, i.e., NUS$_4$, MSRA$_2$ and ANIM$_4$, ACQUEST achieves the runner-up.

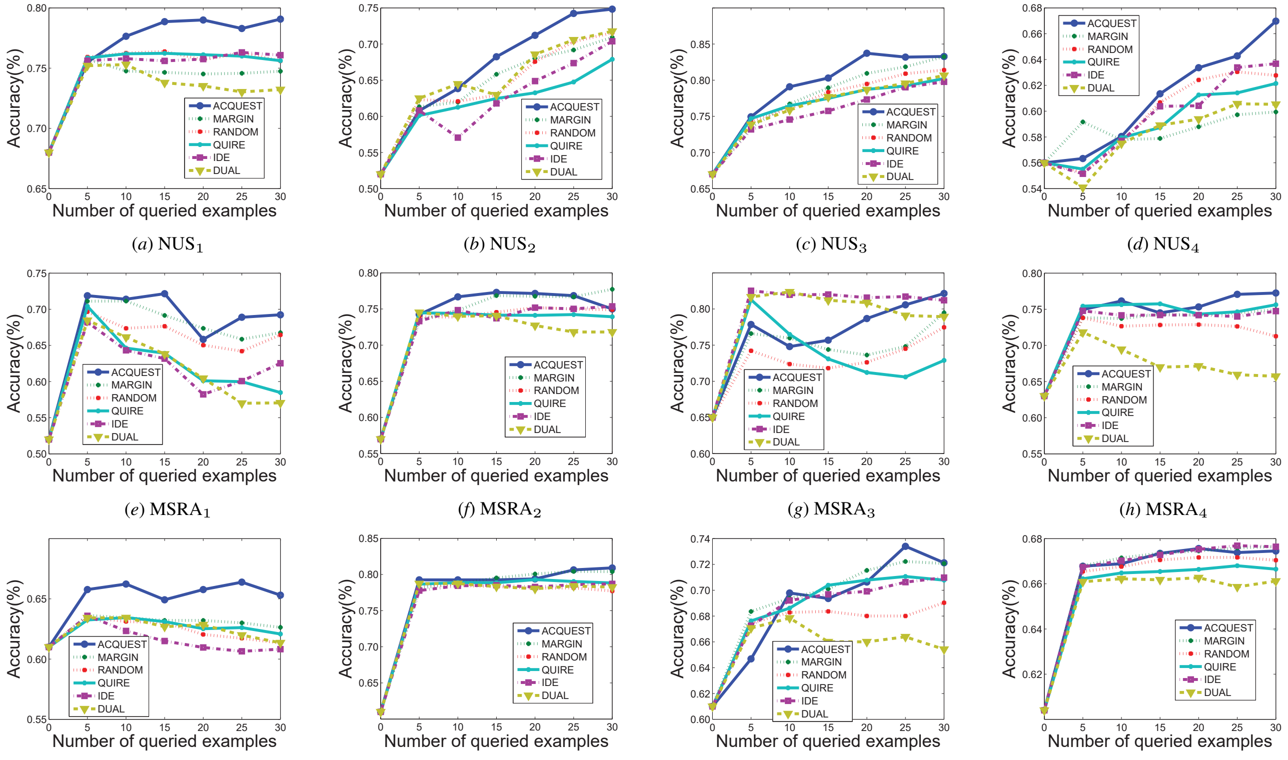
Comparisons when Number of Queries Changes

To investigate the performance of compared active learning methods when number of queries changes, we conduct additional experiments and record more results. Table 2 summarizes the win/tie/loss counts of ACQUEST versus active learning methods with t-test at significance level 95%. The number of queries varies from 5 to 30. From Table 2, the win/tie/lose counts clearly show that ACQUEST approach is superior to those compared active learning methods no matter what the number of queries is.

A detailed classification performance on different number of queries are recorded in the Fig. 1. As in Table 2, the number of queries is set from 5 to 30. From these subplots in Fig. 1, it can be observed that the performance (accuracy) is increased faster for ACQUEST than other compared methods, e.g., on most datasets, ACQUEST gets a high accuracy within 10 queries; besides, the performance of ACQUEST is generally better than other compared approaches on most datasets when the number of queries increases. Moreover, after 10 queries, it can be found that ACQUEST achieves a stable performance on most datasets. Yet the performances of all compared methods are unstable as the queries increasing on some datasets (e.g., MSRA$_1$, MSRA$_3$, ANIM$_1$, ANIM$_3$), this phenomenon can be addressed to the fact that strong modal features, which are picked according to feature extraction costs,

Table 2: Win/tie/loss counts of ACQUEST versus compared methods when the number of queries changes

w/t/l counts	NUMBER OF QUERIES						In All
	5	10	15	20	25	30	
ACQUEST vs MARGIN	7/0/5	10/0/2	10/0/2	10/0/2	11/0/1	10/0/2	58/0/14
ACQUEST vs RANDOM	9/0/3	12/0/0	12/0/0	12/0/0	12/0/0	12/0/0	69/0/3
ACQUEST vs QUIRE	7/0/5	11/0/1	10/0/2	11/0/1	12/0/0	12/0/0	63/0/9
ACQUEST vs IDE	8/0/4	10/0/2	10/0/2	11/0/1	10/0/2	10/0/2	59/0/13
ACQUEST vs DUAL	8/0/4	10/0/2	11/0/1	11/0/1	12/0/0	12/0/0	64/0/8
In All	39/0/21	53/0/7	53/0/7	55/0/5	57/0/3	56/0/4	313/0/47



(*a*) NUS$_1$ (*b*) NUS$_2$ (*c*) NUS$_3$ (*d*) NUS$_4$

(*e*) MSRA$_1$ (*f*) MSRA$_2$ (*g*) MSRA$_3$ (*h*) MSRA$_4$

(*i*) ANIM$_1$ (*j*) ANIM$_2$ (*k*) ANIM$_3$ (*l*) ANIM$_4$

Figure 1: Influence of the query numbers on all compared datasets

are not real "strong" modal features, and perhaps cannot replace the ground truths perfectly for the disturbance of feature noises. Consequently, the pseudo labels from the *inverse prediction* might be less reliable on these datesets.

6 Conclusion

In this work, feature value acquisition expenses for "strong" modal features are considered in both the training and test phase of multi-modal learning. We proposed a new active strong modal feature values acquisition approach called ACQUEST. ACQUEST can exploit the strong modal information by querying the corresponding feature values of selected instances rather than querying the labels directly as in active learning. Due to the fact that feature acquisition hardly needs oracle interventions, ACQUEST requires less human resources, and according to the experiments, our approach achieves comparable or even better performances. It is notable that, different from existing multi-modal learning approaches which take strong/weak modalities in consideration, ACQUEST *actively* selects the instance for strong modal feature acquisition automatically, i.e., ACQUEST can consequently reduce the expenses of strong modal data collecting and labeling in both the training and test phase, while other strong/weak modal learning approaches focus on reducing expenses in test phase. An inverse prediction technique is also proposed for unifying ACQUEST in one optimization formalization as well, which simplifies query selection. Empirical results clearly validate its effectiveness on feature querying, expenses reducing etc. How to identify the real "strong" modality in multi-modal scenarios and extend ACQUEST for multi-class applications should be interesting future works.

References

[Christoph *et al.*, 2009] H.Lampert Christoph, Nichisch Hannes, and Harmeling Stefan. Learning to detect unseen object classes by between-class attribute transfer. In *Proceedings of the 25th Computer Vision and Pattern Recognition*, pages 951–958, Miami, FL, 2009.

[Chua *et al.*, 2009] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yan-Tao Zheng. NUS-WIDE: A real-world web image database from national university of singapore. In *Proceedings of ACM International Conference on Image and Video Retrieval*, pages 1–9, Article No. 48, Santorini, Greece, 2009.

[Fang *et al.*, 2014] Meng Fang, Jie Yin, and Dacheng Tao. Active learning for crowdsourcing using knowledge transfer. In *Proceedings of the 28th Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence*, pages 1809–1815, Quebec, Canada, 2014.

[Huang *et al.*, 2014] Sheng-Jun Huang, Rong Jin, and Zhi-Hua Zhou. Active learning by querying informative and representative examples. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 36(10):1936–1949, 2014.

[Huang *et al.*, 2015] Sheng-Jun Huang, Songcan Chen, and Zhi-Hua Zhou. Multi-label active learning: Query type matters. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence*, pages 946–952, Buenos Aires, Argentina, 2015.

[Ion *et al.*, 2006] Muslea Ion, Minton Steven, and A. Knoblock Craig. Active learning with multiple views. *The Journal of Artificial Intelligence Research*, 27:203–233, 2006.

[Jin *et al.*, 2008] Rong Jin, Jianke Zhu, and Lyu M.R. Semi-supervised SVM batch mode active learning for image retrieval. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1–7, Anchorage, AK, 2008.

[Kiros *et al.*, 2014] Ryan Kiros, Ruslan Salakhutdinov, and Richard Zemel. Multimodal neural language models. In *Proceedings of the 31st International Conference on Machine Learning*, pages 595–603, Beijing, China, 2014.

[Maria-Florina *et al.*, 2007] Balcan Maria-Florina, Broder Andrei, and Tong Zhang. Margin based active learning. In *Proceedings of the 20th Annual Conference on Learning Theory*, pages 35–50, 2007.

[Nguyen *et al.*, 2013] Cam-Tu Nguyen, De-Chuan Zhan, and Zhi-Hua Zhou. Multi-modal image annotation with multi-instance multi-label LDA. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence*, pages 1558–1564, Beijing, China, 2013.

[Pinar *et al.*, 2007] Donmez Pinar, Carbonell Jaime G., and Bennett Paul N. Dual strategy active learning. In *Proceedings of the 18th European Conference on Machine Learning*, pages 116–127, SWarsaw, Poland, 2007.

[Sachan *et al.*, 2015] Mrinmaya Sachan, Eduard Hovy, and Eric P. Xing. An active learning approach to coreference resolution. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence*, pages 1312–1318, Buenos Aires, Argentina, 2015.

[Sanjoy and Daniel, 2008] Dasgupta Sanjoy and Hsu Daniel. Hierarchical sampling for active learning. In *Proceedings of the 25th International Conference on Machine Learning*, pages 208–215, Helsinki, Finland, 2008.

[Settles, 2010] Burr Settles. Active learning literature survey. *University of Wisconsin, Madison*, 52(55-66):11, 2010.

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

[Wang *et al.*, 2009] Meng Wang, Linjun Yang, and Xian-Sheng Hua. MSRA-MM: Bridging research and industrial societies for multimedia information retrieval. Technical report, Microsoft Research Asia, Microsoft, 2009.

[Wang *et al.*, 2015] Daixin Wang, Peng Cui, Mingdong Ou, and Wenwu Zhu. Deep multimodal hashing with orthogonal regularization. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence*, pages 2291–2297, Buenos Aires, Argentina, 2015.

[Yang *et al.*, 2015] Yang Yang, Han-Jia Ye, De-Chuan Zhan, and Yuan Jiang. Auxiliary information regularized machine for multiple modality feature learning. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence*, pages 1033–1039, Buenos Aires, Argentina, 2015.

[Zhang and Li, 2014] Dongqing Zhang and Wu-Jun Li. Large-Scale supervised multimodal hashing with semantic correlation maximization. In *Proceedings of the 28th Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence*, pages 2177–2183, Quebec, Canada, 2014.

[Zhang *et al.*, 2014] Qing Zhang, Yilong Yin, De-Chuan Zhan, and Jingliang Peng. A novel serial multimodal biometrics framework based on semi-supervised learning techniques. *IEEE Transactions on Information Forensics and Security*, 9(10):1681–1694, 2014.

[Zhong *et al.*, 2015] Jinhong Zhong, Ke Tang, and Zhi-Hua Zhou. Active learning from crowds with unsure option. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence*, pages 1061–1067, Buenos Aires, Argentina, 2015.

[Zhou *et al.*, 2005] Xiaoli Zhou, Bir Bhanu, and Ju Han. Human recognition at a distance in video by integrating face profile and gait. *IEEE Transactions on Face Biometrics for Personal Identification*, 3546(5):533–543, 2005.

[Zhou *et al.*, 2015] Dawei Zhou, Jiebo Luo, Vincent Silenzio, Yun Zhou, Jile Hu, Glenn Currier, and Henry Kautz. Tackling mental health by integrating unobtrusive multimodal sensing. In *Proceedings of the 29th AAAI*, pages 1401–1408, Austin, Texas, 2015.