Instance Specific Discriminative Modal Pursuit: A Serialized Approach*

Yang Yang YANGY@LAMDA.NJU.EDU.CN
De-Chuan Zhan ZHANDC@LAMDA.NJU.EDU.CN
Ying Fan FANY@LAMDA.NJU.EDU.CN
Yuan Jiang JIANGY@LAMDA.NJU.EDU.CN
National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China

Editors: Yung-Kyun Noh and Min-Ling Zhang

Abstract

With the fast development of data collection techniques, a huge amount of complex multi-modal data are generated, shared and stored on the Internet. The burden of extracting multi-modal features for test instances in data analysis becomes the main fact that hurts the efficiency of prediction. In this paper, in order to reduce the modal extraction cost in serialized classification system, we propose a novel end-to-end serialized adaptive decision approach named Discriminative Modal Pursuit (DMP), which can automatically extract instance-specifically discriminative modal sequence for reducing the cost of feature extraction in the test phase. Rather than jointly optimize a highly non-convex empirical risk minimization problem, we are inspired by LSTM, and the proposed DMP can turn to learn the decision policies which predict the label information and decide the modalities to be extracted simultaneously within limited modal acquisition budget. Consequently, DMP approach can balance the classification performance and modal feature extraction cost by utilizing different modalities for different test instances. Empirical studies show that DMP is more efficient and effective than existing modal/feature extraction methods.

Keywords: Multi-Modal Learning; Serialized Modal/Feature Extraction

1. Introduction

With the rapid development of data collection techniques, complicated objects can always be represented as multi-modal features, and it becomes a great challenge to design efficient and effective methods on these accumulated multi-modal data. Therefore, efficient online information processing approaches are urgently needed. Researchers have developed lots of online learning methods, such as online clustering (Gentile et al., 2014), online classification (Babenko et al., 2011), and try to speed up the learning by sampling or advanced optimization techniques (Zhang et al., 2016), these methods always focus on instances level online problem. However, different modalities are with various extraction expenses, and previous researches, i.e., dimensionality reduction methods, generally assume that all the multi-modal features of test instances have been already extracted without considering the overall costs. While it is usually the case that for unseen test instances, there is no aforehand multi-modal features prepared, modal extraction need to be performed in the test phase at

* This work was supported by NSFC (61632004, 61773198), Huawei Fund (YBN2017030027) and the Collaborative Innovation Center of Novel Software Technology and Industrialization.

first. While for the complex multi-modal data collection nowadays, the heavy computation burden of feature extraction for different modalities has become the dominant fact that hurts the efficiency.

Traditional modal extraction or feature extraction approaches originated from cascade detection (Viola and Jones, 2001), which try to rank all modalities or features in a descending rank according to the discriminative power or in an ascending list according to the extraction cost firstly, then perform modal/feature extraction with fixed acquisition budget or specified precision. These methods, in fact, fixed the modal/feature extraction order for all instances. However, in real applications, different instances should have different modal/feature extraction order, in detail, the next modality to be extracted should be decided by the value of those extracted modal features, e.g., aiming at a particular disease as Fig. 1 shows, after common examination for all patients, different patients will perform different subsequent examinations according to the results of the common examination. In other words, to predict an unseen instance as efficient and accurate as possible, we need to figure out a modal extraction for the concerned instance individually. For the absence of discriminative modal feature values on the unseen instances, it is supposed that for all test instances, there are a same "initial modality", which can be the most likely to be available in raw data, e.g., the common examination of majority disease, the "initial modality" can be the blood routine tests; for image data, the "initial modality" can be the pixels values (RGB or gray levels).

Aiming at classifying typical instances with inexpensive modalities while using expensive modalities for atypical instances, this is inspired by adaptive decision methods for feature extraction, and reduced the overall costs in return (Kusner et al., 2014; Kanani and Melville, 2008), which construct a tree of classifiers to reduce the average test time complexity of test instances, while maximize their accuracy. Specially, it is a fact that several modalities usually can be sufficient for predicting some "easy" instances with high accuracy, e.g., the qualified face images can be detected accurately with some "simple but powerful" face template features. Thus, for this kind of "easy" test instances, it is not necessary to extract other kinds of features, while we extra expensive features for some "hard" instances. This phenomenon suggests spending different efforts of feature extraction on different test instances to reduce the costs. However, these feature extraction methods are always non-convex problems to solve, and are difficult to handle the multi-modal data, which have more practical significance in real application, e.g., we always extract multiple feature subsets as the blood test shown in Fig. 4 rather than single feature for classification or other tasks, thus several adaptive decision methods for modal extraction are proposed (Wang et al., 2014, 2015). Nevertheless, previous directed acyclic graph based methods need to list all the enumeration of probable modal sequences, which will be combination explosion with large number of modalities.

To solve these problems, in this paper, we proposed a novel end-to-end network DMP (Discriminative Modal Pursuit) approach for serialized modal feature extraction, different from previous feature selection or dimensionality reduction methods, DMP extracts different modalities for different instances, and Fig. 2 gives a difference illustration on the number of features invoked during testing for DMP, feature selection, from which it can be found that our approach can further reduce the number of features for testing for the nature of "easy" instances existing. Besides, DMP handles modalities sequentially, which is more efficient and accurate than previous modal/feature extraction methods. DMP is inspired by LSTM network, which can predict label and following modality simultaneously with a fixed budget, specifically, when a test instance is received, DMP firstly extract the same "initial modality", then if DMP is confident to assign the class label, classification will be made, otherwise, DMP will decide which modality should be extracted next, this procedure will repeat until the cost budget is overstepped or it can make a confident classification.

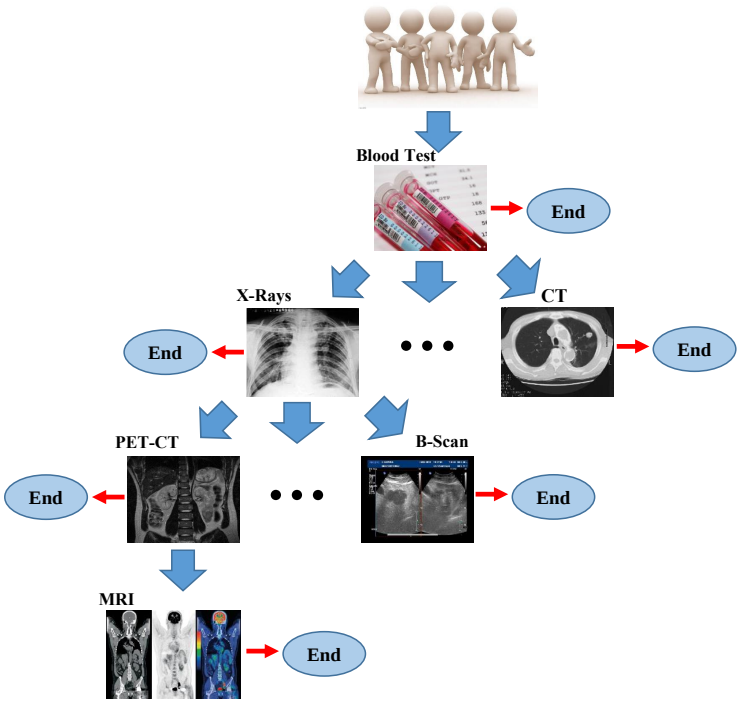


Figure 1: The overall flowchart. For the same disease, patients are always checked with the same examination first, then partial patients will take different consecutive examinations with various costs based on the previous examinations, while some "easy" diagnostic patients can end the examination with high accuracy. The procedure aims at making a convince prediction with least overall costs.

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

The problem of modal extraction for reducing the overall cost has been extensively researched. Originally, feature selection and dimensionality reduction are generally used for reducing the expenses of feature extraction and can be adapted to modal extraction problems. Song and Lu (2017) proposed Regularized multilinear regression and selection for automatically selecting a subset of features while optimizing prediction for multidimensional data; Jian et al. (2016) used a novel multi-label informed feature selection framework MIFS, which exploits label correlations to select discriminative features across multiple labels. However, as shown in Fig. 2, feature selection methods are to select a subgroup of features which are consistently used during the whole test phase. While the problem to tackle in this paper is to reduce the test feature extraction expenses by choosing different subgroups of features for different instances, and further reduce the feature extraction costs from modalities. In other words, the goal of this work is aiming at extraction instance specific modal features in test phase. As for dimensionality reduction methods, the discriminative informa-

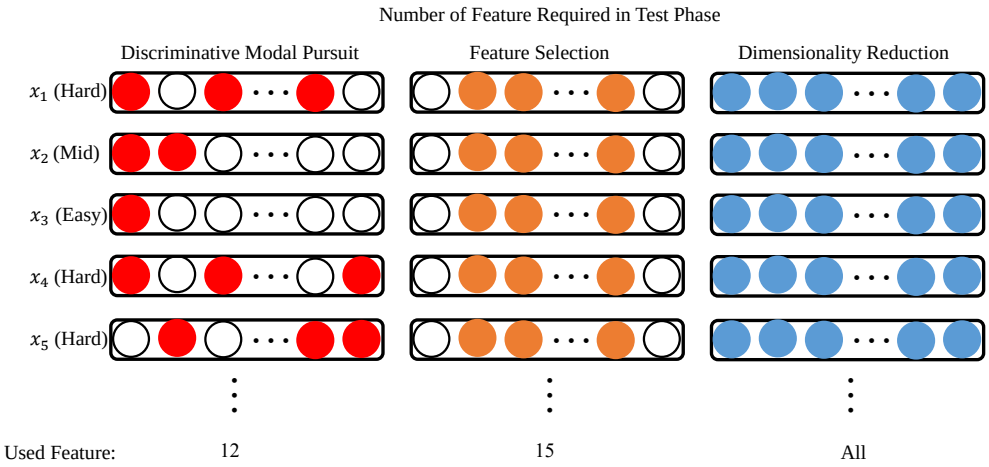


Figure 2: The difference of the number of features used in test phase among Discriminative Modal Pursuit, feature selection, and dimensionality reduction. In the plots above circles shadowed with colors represent the features used (different colors are corresponding to different methods), and x_1, x_4, x_5 are "hard" instances which may consume more feature values for good prediction while x_3 is an "easy" one that only need features from 1 modality. It can be found that feature selection and dimensionality reduction methods use the same features are all instance, especially for dimensionality reduction, the subspace learned can be extracted from almost all raw features, i.e., dimensionality reduction methods require the most feature value extraction costs. While on the other hand, Discriminative Modal Pursuit adopt different features for different instance, and as a consequence, can reduce the overall feature extraction cost.

tion are emphasized, and as a matter of fact, almost the entire feature sets from all modalities are required in a large portion of dimensionality reduction methods. Therefore, the target differences between this work and general dimensionality reduction are even larger.

Thus, Chen et al. (2012); Zhang et al. (2014); Liu et al. (2008) proposed cascade detection methods to reduce the overall costs, which first compute the cheap features for filtering out the negative examples and more expensive are acquired later, however, these approaches require the fixed acquisition order of features and generalize to multi-class difficultly. Then, Gao and Koller (2011); Panella and Gmytrasiewicz (2016) posed the system as a PO-MDP problem with generative methods by selecting modalities based information gain of unknown modalities, while they always require estimation of the probability distributions. To overcome this problem, Karayev et al. (2013) used imitation learning approaches, which turn to predict the reward or oracle action, nevertheless, these methods require operate on a wide range of missing feature patterns.

Consequently, in order to learn efficient and effective instance specific modal/feature extraction model for cost reduction, Xu et al. (2013) introduced cost-sensitive trees of classifies (CSTC), CSTC is a global empirical risk minimization problem using tree structure, which applied internal decision rules and leaf classifiers, jointly with alternative minimization techniques; Kusner et al. (2014) proposed approximately sub-modular trees of classifiers ASTC, a variation of CSTC which employs

a different relaxation using approximate sub-modularity, and significantly reduced the training time; Nan et al. (2015) proposed random forests to efficiently learn a budgeted decision rules, however, they are all the feature extraction methods, which always optimize and expand to modal extraction difficult, recently, Wang et al. (2015) proposed an adaptive sensor acquisition system, which is modeled as a directed acyclic graph (DAG) for resource constrained prediction. Nevertheless, it needs to list all permutations of modal sequence in the training phase.

In this paper, we proposed the DMP (Discriminative Modal Pursuit) approach for learning a serialized modal extraction decision methods inspired from LSTM. Specially, with the given raw multi-modal training data, we learn an instance specific discriminative modal extraction method. It is required to answer two closed related questions, i.e., which modality should be extracted next? when to stop extracting new modality and make a prediction? Consequently, when presented with an unseen instance, we would extract the most informative and cost-effective modal sequence for the instance. Empirical study shows the efficiencies and effectiveness of DMP, i.e., it is able to reduce the modal extraction cost and achieve even better classification performance.

3. Proposed Method

This section mainly gives the detail description of the Instance Specific Discriminative Modal Pursuit (DMP) approach after a preliminary notation explanation.

3.1. Notation

Suppose there are N labeled examples for training, which are denoted as $\mathcal{T} = \{\mathbf{x}_i, y_i\}_{i=1}^{N}$, where $\mathbf{x}_i \in \mathbb{R}^d$, $y_i \in \{1, 2, \cdots, k\}$. In order to facilitate the discussion, we represent the example $\mathbf{x}_i$ as $\{\mathbf{x}_i^1, \mathbf{x}_i^2, \cdots, \mathbf{x}_i^M\}$, M is the number of modality, where $\mathbf{x}_i^j$ is j-th modality of $\mathbf{x}_i$ with d_j dimension, thus, there are M disjoint modal features for each examples and the cardinality of the union of M modalities is d. Meanwhile, different modalities are with various costs, and we denoted the costs of the modalities as $\mathcal{C} = \{c^1, c^2, \cdots, c^M\}$. Meanwhile, the test instance is denoted as X_t, which only given the "initial modality".

Here we consider there are no repeat modalities selected in the same modal extraction sequence for specific instance. We denote the serial number of modality corresponding to the j-th decision step of $i-$th instance as $s_i^j \in \mathcal{S}, 1 \leq i \leq N, 1 \leq j \leq M$, where $\mathcal{S} = \{1, 2, \cdots, M\}$ is the set of all possible values of s_i^j, i.e., the j-th step selected modality of instance i is $\mathbf{x}_i(s_i^j)$, we can denote the relationship between $\mathbf{x}_i$, s_i^j and $\mathbf{x}_i^m$ as $\mathbf{x}_i(s_i^j) = \mathbf{x}_i^m$, where $s_i^j = m$. Note that the modal features of s_i^j could be different from that of s_k^j, $i \neq k$ and $j \neq 1$, because even the same modality can have different discriminative power on different instances, thus for the i-th instance, we should extract, e.g., modality A on the j-th step and for the k-th instance, modality B of features should be extracted instead on the step j, i.e., the feature subset extracting order is heavily depend on specific instances, while the lengths of modal extraction sequence for different instances are also different.

3.2. Discriminative Modal Pursuit

In this section, the original problem can be considered as a degenerated case, e.g., as shown in Fig. 1 when the patients are checked the same disease, it usually performs the common examination for every patients first, i.e., blood test. Then, different patients will take different consecutive examinations, i.e., CT, X-Rays, PET-CT etc, which are referred to the results of current examination, while

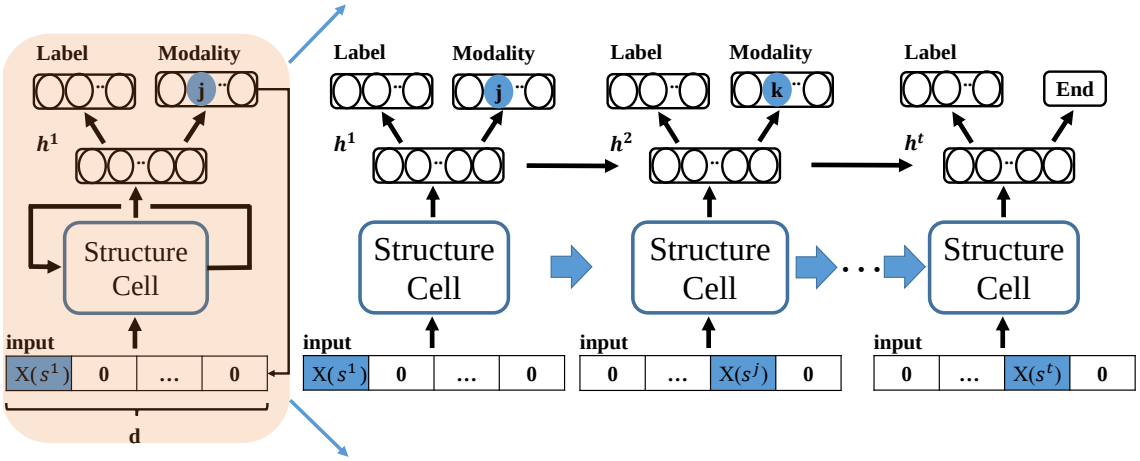


Figure 3: The overall flowchart of DMP approach, right part is the decomposition of the left part. As shown in the schematic, the "initial modality" is same for both training and test data. And at $j-$th step, the input modal features are $\mathbf{x}(\mathbf{s}^j)$, shown in blue shadows, it is notable that different instances can with various input at the same step, and different instances may with various length of modal extraction sequence. Then we input the modified $x(s^j)$ to the DMP network. With the calculated output features $\mathbf{h}$, we predict the label information and following modality as the next step input simultaneously. We repeat the procedure until reach the end criterion.

some "easy" diagnostic patients can end the examination with high accuracy, therefore, there are particular examination sequence for different patients with least overall cost and high accuracy. In other words, in our setting, given the complete disordered multi-modal training examples, we need to efficiently learn an adaptive modal extraction decision method for overall cost reduction, while maximize the accuracy.

In this situation, the training examples can be denoted as $\mathcal{T} = \{\mathbf{x}_i, y_i\}_{i=1}^n$, which are with complete disordered multi-modal features, and test instance is denoted as X_t, Note that only the modality s^1 is defined for both the training and test data, while there are no more modality provided for test instances. For simplicity, the $\mathbf{x}_i(s^1)$ is the "initial modality", which can be obtained easily in applications, e.g., the pixels of raw images, the blood test in disease examinations etc. Based on the "initial modality", in order to classify the test instance more efficiently and effectively, there are two learning tasks should be performed as following:

- Which modality should be extracted next?
- When to stop extracting new modality and make a prediction?

Modal Extraction Decision:
The first question indicates that to classify an unseen test instance efficiently, modalities with more discriminative abilities should be extracted if previous modalities are difficult to make a convince prediction, otherwise, some "easy" modalities can be selected. To answer the question, it is needed to design an adaptive modal extract decision method to automatically extract the next informative

and discriminative modality based on aforementioned modalities. Because even the same modality can have different discriminative power for different instances, the next modality to be extracted should be related to the instance's extracted modal features.

Without any loss of generality, we denote the modal sequence of $i-$th instance to be learned as $\mathbf{s}_i = \{s_i^1, s_i^2, \cdots, s_i^k\}, k \leq M$, k is different for different instances. It obviously indicates that for the i-instance, when the initial modality s_i^1 is given, then we will extract s_i^2 modality and so forth. At the t-th step, the "modal extraction sub-sequence" can be denoted as $\mathbf{s}_i^t = \{s_i^1, s_i^2, \cdots, s_i^t\} \in \mathcal{X}, t \leq k$, where $\mathcal{X}$ is the power set of all possible sequences of modal feature values extracted, and the modality to be extracted in next step is s_i^{t+1}. The first question now can be rewritten as "given the extracted modal sequence $\mathbf{s}_i^t$, how to decide the next modality s_i^{t+1} to be extracted". The question can be further formalized to learn an adaptive decision map function: $\mathbf{G} : \mathcal{X} \rightarrow \hat{\mathcal{S}}$, where $\hat{\mathcal{S}}$ is the set consisted of all possible values of $s_i^j \in \{\mathcal{S} - \mathbf{s}_i^t\}$, which means there have no repeat modality in the same modal extraction sequence.

As a matter of fact, the goal of learning adaptive decision map $\mathbf{G}$ needs to utilize the previous extracted modal features. We put forward a novel end-to-end deep network structure DMP inspired from LSTM network (Graves and Schmidhuber, 2005), previous LSTM network used purpose-built memory cells to store information, which can be considered as the previous selected modalities, and is better at finding and exploiting long range information. It is notable that the input of LSTM orients to the corresponding data of the sequence in each step, thus the map function $\mathbf{G}$ is equivalent to $\mathcal{H} \rightarrow \hat{\mathcal{S}}$, where $\mathcal{H}$ is the space of the output in each step. In each step of DMP, the input is the extracted modal features of last step, then regard the predicted modality as the input for next step, these procedures repeat until reaching the stop criterion. The overall flowchart of DMP approach is shown in Fig. 3, the right part is the spread of left part in shadow.

In detail, firstly, with the multi-modal training examples $\mathcal{T}$, considering the dimensions of different modalities is heterogeneous, which is difficult input to the network. Thus, as shown in the bottom of left part in Fig. 3, the extracted modality at $t-$th step can be represented as s^t, and the raw input modal features at $t-$th step can be denoted as $\mathbf{x}(s^t)$. Then, we modify the raw input $\mathbf{x}(s^t) \in \mathbb{R}^{d_{s^t}}$ to $\hat{\mathbf{x}}(s^t) = [\hat{\mathbf{x}}^1, \hat{\mathbf{x}}^2, \cdots, \hat{\mathbf{x}}^M] \in \mathbb{R}^d$, in which $\hat{\mathbf{x}}^j$ has real values if $j = s^t$, otherwise, $\hat{\mathbf{x}}^j = 0$. And we denote the $\hat{\mathbf{x}}(s^t)$ as the new input of the network. Finally, the output features of DMP network can be represented as $\mathbf{h}^t \in \mathbb{R}^h$ by calculating among several layers by the following equation:

$$
\begin{aligned}
i^t &= \sigma(W_{xi}\hat{\mathbf{x}}(s^t) + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \\
f^t &= \sigma(W_{xf}\hat{\mathbf{x}}(s^t) + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \\
c^t &= f_t c_{t-1} + i_t \tanh(W_{xc}\mathbf{x}(s^{(t-1)}) + W_{hc}h_{t-1} + b_c) \\
o^t &= \sigma(W_{ox}\hat{\mathbf{x}}(s^t) + W_{ho}h_{t-1} + W_{co}c^t + b_o) \\
\mathbf{h}^t &= o_t \tanh(c_t)
\end{aligned}
$$

where the weight matrices from the cell to gate vectors (e.g. W_{xi}) are diagonal, the b terms denote bias vectors (e.g. b_i is hidden bias vector), and i, f, c and o are respectively the *input gate*, *forget gate*, *output gate* and *cell activation vectors*, all of which are the same size as the hidden vector $\mathbf{h}$. And all the parameter facts in network can be denoted as Θ.

Then, at the $t-$th step, we want to make the label prediction and modal prediction simultaneously with all the previous outputs $\{\mathbf{h}^1, \mathbf{h}^2, \cdots, \mathbf{h}^{t-1}\}$. Thus, we directly stack the whole previous outputs as $\hat{\mathbf{h}}^t = [\mathbf{h}^1, \mathbf{h}^2, \cdots, \mathbf{h}^{t-1}] \in \mathbb{R}^{(t-1)h}$, and arrange the ground truth prediction and modal

prediction in parallel to the stacked output features $\hat{\mathbf{h}}^t$. The fully connected weights between $\hat{\mathbf{h}}^t$ and label concepts can be organized as a linear mapping matrix V^t together with a nonlinear softmax function. Besides, in this parallel structure, there are also linear connections between the output features and the modal prediction layer, which is also a full connection structure, these weights can be denoted as U^t. It is notable that there are independent mapping matrix V and U in each step and maximum M modalities can be extracted in a sequence.

As a general training procedure, it focuses on reducing the errors made in the current status of the network. Considering there are two predict targets in parallel in the network as shown in the top of left part in Fig. 3. Without any loss of generalities, the loss function implied in the parallel network structure is:

$$L(\Theta, V^1, V^2, \cdots, V^M, U^1, U^2, \cdots, U^M) = \sum_{i=1}^{N} \sum_{j=\mathbf{s}_i^1}^{\mathbf{s}_i^M} (\ell(f(\mathbf{x}_i(j)), \mathbf{G}(\mathbf{x}_i(j)), y_i) + \lambda L_{reg}), \quad (1)$$

Where $\ell(f(\mathbf{x}_i(j)), \mathbf{G}(\mathbf{x}_i(j)), y_i)$ is the loss function of label prediction and modal prediction, L_{reg} is the regularization of the parameters. Note that each step is independent with each other, and we can divide the Eq. 1 into M subproblem, M is the number of modalities. Meanwhile, different instances may have various lengths of modal sequence, in other words, the loss of $i-$th instance will be 0 if it has stopped at $t-$th step. Thus, at $t-$th step, the loss function can be represented as:

$$L^t(\Theta, V^t, U^t) = \sum_{i=1}^{N} (\ell(f(\mathbf{x}_i(s^t)), \mathbf{G}(\mathbf{x}_i(s^t)), y_i) + \lambda L_{reg}^t), \quad (2)$$

where

$$\begin{aligned}
\ell(f(\mathbf{x}_i(s^t)), \mathbf{G}(\mathbf{x}_i(s^t)), y_i) &= \widetilde{\ell}(f(\hat{\mathbf{h}}_i(s^t), y_i)) + \hat{\ell}(\hat{\mathbf{x}}_i(s^t), \mathbf{G}(\hat{\mathbf{h}}_i(s^t)));\\
\widetilde{\ell}(f(\hat{\mathbf{h}}_i(s^t)), y_i) &= y_i \log g(f(\hat{\mathbf{h}}_i(s^t)));\\
\hat{\ell}(\hat{\mathbf{x}}_i(s^t), \mathbf{G}(\hat{\mathbf{h}}_i(s^t))) &= -\frac{1}{2}\|\hat{X}_i(s^t)\mathbf{G}(\hat{\mathbf{h}}_i(s^t)) - \hat{\mathbf{x}}_i(s^t)\|_F^2 (1 - \widetilde{\ell}(f(\hat{\mathbf{h}}_i(s^t)), y_i)).
\end{aligned}$$

Here $\hat{\mathbf{x}}_i(s^t)$ is the transformation of the raw input $\mathbf{x}_i(s^t)$, $\hat{\mathbf{h}}_i(s^t)$ is the stacked output features of $t-$th step with input $\hat{\mathbf{x}}_i(s^t)$, $\widetilde{\ell}(f(\hat{\mathbf{h}}_i(s^t)), y_i)$ is the label prediction loss function ($\widetilde{\ell}$ can be with any convex loss functions), in which $f(\hat{\mathbf{h}}_i(s^t))$ is the prediction of $\hat{\mathbf{h}}_i(s^t)$, we define as linear function $\hat{\mathbf{h}}_i(s^t)V^t + b_{V^t}$ for simplicity here, b_{V^t} is the bias for predictors of $\hat{\mathbf{h}}_i(s^t)$, g is a softmax operator. It it notable that $\mathbf{G}(\hat{\mathbf{h}}_i(s^t)) \in \hat{\mathcal{S}}$ is the modal prediction of $\hat{\mathbf{h}}_i(s^t)$, we also define as linear function $\hat{\mathbf{h}}_i(s^t)U^t + b_{U^t}$, b_{U^t} is the bias for predictors U^t. Finally, we select the modality with $\max \mathbf{G}(\hat{\mathbf{h}}_i(s^t))$ prediction as the result.

The main targets of DMP are closely related to the $\hat{\ell}(\hat{\mathbf{x}}_i(s^t), \mathbf{G}(\hat{\mathbf{h}}_i(s^t)))$, which is the loss of modal prediction, as a matter of fact, the raw data are disordered, thus there have no ground truth for modal prediction, however, the main fact of deciding which modality need to be extracted with priority is the current modal importance, which generally represent by the classification performance: $1 - \widetilde{\ell}(f(\hat{\mathbf{h}}_i(s^t)), y_i)$, reciting that more discriminative modalities should be extracted if current modality is weak, thus, modal discriminative ability should also be considered, many discriminative ability measures are related to the distances within neighborhoods, hence $\hat{X}_i(s^t)\mathbf{G}(\hat{\mathbf{h}}_i(s^t)) - \hat{\mathbf{x}}_i(s^t)$ measures the distance between prediction modal and current modality, where $\hat{X}_i(s^t) \in \mathbb{R}^{d \times M}$ is

Algorithm 1 Training Algorithm For DMP

Input
- $\mathcal{T} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, $\mathcal{C} = \{c^1, c^2, \cdots, c^M\}$: Training data with raw multi-modal features and costs; max-iter: k; Budget: cost budget $\mathrm{c_{thr}}$ and prediction confidence threshold $\mathrm{a_{thr}}$.

Output
- $\mathbf{G}$: Learner to predict which subset of features should be extracted in next step
- f: Classifier trained with acquired sequence modal features

Steps:

Modify the raw multi-modal data $\mathbf{x}_i^j \in \mathcal{R}^{d_j}$ to $\hat{\mathbf{x}}_i^j \in \mathcal{R}^d$ with the same dimension

repeat

  Create Batch: Randomly pick up n examples from $\mathcal{T}$ without replacement

  Input the "initial modality" $\hat{\mathbf{x}}_i(s^1), i = 1, 2, \cdots, n$

  for $t = 1, 2, \cdots, M-1$ **do**

    if $\hat{\mathrm{c}}_{\mathrm{i}}^{\mathrm{t}} \leq \mathrm{c_{thr}}$ and $\mathrm{a}_{\mathrm{i}}^{\mathrm{t}} \leq \mathrm{a_{thr}}$ **then**

      Calculate the loss L^t in Eq. 2;

      Weight Propagation step: Obtain the derivative $\partial L^t/\partial V^t$, $\partial L^t/\partial U^t$, $\partial L^t/\partial \Theta$;

      Update parameters V^t, U^t, Θ;

      Extract the modality with $\max \mathbf{G}(\hat{\mathbf{h}}_i(s^t))$ as the next input;

    else

      Define the $L_i^t = 0$;

    end if

  end for

until converge or reach the max-iter: k

the combination of the transformed prediction modal feature values, each column corresponds a prediction modality. In general, we consider the modal importance is related with the prediction performance, and we wish to extract the modalities that are always from current modality if the importance is weak, which will be more discriminative. By simultaneously considering the modal specific property and the modal power between input modal features and neighborhoods. L_{reg}^t can be any convex regularization, in order to facilitate selecting the most discriminative modality with least overall cost, in this paper, we choose L_{reg}^t as: $\|V\|_2^2 + \|U\|_2^2 + \|C\mathbf{G}(\mathbf{h}_i(s^t))\|_2^2$, where C is the vector of modal cost corresponding to the predict modalities of $\mathbf{G}(\mathbf{h}_i(s^t))$, note that there have no repeated modalities in the same modal sequence, so we will set the c of the extracted modalities before large enough. The parameter λ controls the trade-off between the loss and regularization. In the training procedure, the derivatives are taken with the help of back propagation technique. The detail training procedure is shown in Alg. 1.

Stop Criterion:

It is notable that there is no need to obtain the full modal extraction sequence for each test instance, because instances can be predicted accurately with a subset of modalities in most case. In fact, our second question also implies that to efficiently classify an unseen test instance, we need to define a stopping criterion for our framework. When the modal extraction process reaches the criterion, further extracting of new modalities will be stopped and the label of the concerned instance will be predicted based on the extracted modalities.

Two kinds of stopping criteria are selected in this work. The first one is the modal extracting cost budget, c_{thr} provided by users, and we define the accumulated cost at $t-$th step as $\hat{c}^t = c(s^1) + c(s^2) + \cdots + c(s^t)$. Another kind of stopping criterion is based on the confidence of the prediction for the instance, which means the DMP choose to stop, and we define the prediction confidence at $t-$th step as $\hat{a}^t$.

Prediction Procedure:
In the test phase, all the instance are only given the "initial modality" $x(s^1)$ and the cost budget c_b, with the trained decision learner G and label prediction classifier f, we input "initial modality" $x(s^1)$, then run the instance through the learner G and obtained the modal extraction sequence $\mathbf{s}$ with final prediction label from f.

4. Experiments

In the serialized modal extraction problem, DMP can classify new instances only with the "initial modality" of least cost. In this section, we will provide the empirical investigations and performances of DMP. In particular, investigations on 2 real world small datasets and 5 large datasets, i.e., letter classification, satellite classification, the electron neutrinos classification, forest cover classification, image classification (Cifar, NUS, Scene).

In 3 small datasets, we use the multi-modal features given by the raw data. While in 3 large datasets, there are no explicit modal partitions, and we calculate the information entropy gain for each features, then divide the features into different modalities as the same number in (Wang et al., 2015). For 3 small datasets and NUS, Scene datasets, 66% instances are chosen as training set and the remains are test set. In other three large datasets, training and test splits are provided by (Wang et al., 2015), i.e., there are 45,523/19,510/65,031 examples in training/validation/test sets for MiniBooNE dataset, 36,603/15,688/58,101 examples in training/validation/test sets for Forest dataset, 19,761/8,468/10,000 examples in training/validation/test sets for Cifar-10 dataset. We repeat experiments 30 times on each dataset, the average error rates are recorded and evaluated. The parameter λ in the training phase is tuned in $\{0.1, 0.2, \cdots, 0.9\}$. When the variations between the objective value of Eq. 1 is less than 10^{-5} in iterations, we consider DMP converges. We run the following experiments with the implementation of an environment on NVIDIA K80 GPUs server and our model can be trained about 290 images per second with a single K80 GPU.

Some of the previous sequence modal extraction methods can be used, thus, DMP is compared to the novel used modal extraction method, i.e., LP, DAG. For DMP can also be degenerated to feature extraction approaches, CSTC, ASTC, Greedy Miser are also compared in our experiments. In detail, the compared methods are listed as:

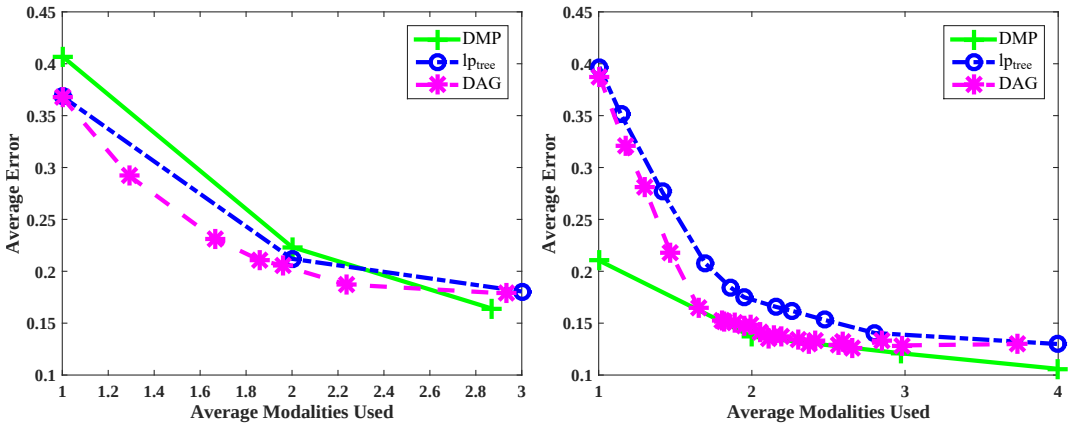
- **Lptree**: first applies the LP approach to learning the modal trees as (Joseph Wang and Saligrama, 2014), and then construct trees containing all subsets of sensors as opposed to fixed order cascades (Wang et al., 2015);

- **DAG**: proposes an adaptive sensor acquisition system modeled as a directed acyclic graph, where sensors can be regard as modalities (Wang et al., 2015);

- **Greedy Miser**: incorporates the feature extraction cost during training to explicitly minimize the cpu-time during testing, which is a straightforward extension of stage-wise regression and is equally suitable for regression or multi-class classification (Xu et al., 2012);

Table 1: The avg. error rates and feature extraction cost with compared methods of small datasets. The significant best classification performance on each dataset is bolded. Extraction cost is defined as the average modal/feature used of all test instances.

	DMP	Lptree	DAG	DMP	Lptree	DAG
	Average Error Rates (mean±std.)			Extraction Cost (mean±std.)		
Letter	**.164±.031**	.180±.032	.179±.024	**2.867±.046**	3.000±.000	2.931±.013
Landsat	**.106±.020**	.130±.013	.130±.006	3.996±.007	4.000±.000	**3.735±.045**



(*a*) Letter (*b*) Landsat

Figure 4: Average number of modalities acquired vs. test error comparison.

- **CSTC**: is a global empirical risk minimization problem using tree structure, which applies internal decision rules and leaf classifiers, jointly with alternative minimization techniques (Xu et al., 2013);
- **ASTC**: proposes approximately submodular trees of classifiers, a variation of CSTC which employs a different relaxation using approximate submodularity (Kusner et al., 2014).

4.1. Small Datasets

In small datasets, 2 previous datasets for budget cascades are tested: the Letter dataset consists of modalities extracted from hand written digits, i.e., the first modality is with five features generated from position and pixel counts, the next modality is with 7 features in the second stage correspond to more complex features such as spatial moments, and the final modality consists 4 features in stage 3 correspond to the most complex features, such as edge based features. Landsat data consists of 3×3 pixel neighborhoods taken from a satellite image at four different hyper spectral bands, specifically, the four spectral values for the top-left pixel are given first followed by the four spectral values for the top-middle pixel and then those for the top-right pixel, and so on with the pixels read out in sequence left-to-right and top-to-bottom. Thus, the four spectral values for the central pixel

Table 2: The avg. error rates and feature extraction cost with compared methods of large datasets. The significant best classification performance on each dataset is bolded. Extraction cost is defined as the average modal/feature used of all test instances.

	DMP	CSTC	ASTC	Greedy Miser	DAG
	Average Error Rates (mean±std.)				
MiniBooNE	.082±.042	.160±.035	.103±.027	**.073±.055**	.089±.045
Forest	**.137±.029**	.249±.044	.236±.024	.144±.038	.213±.030
Cifar	**.291±.013**	.335±.025	.339±.035	.295±.067	.315±.047
NUS	**.351±.006**	.357±.016	N/A	.361±.038	N/A
Scene	**.297±.046**	.317±.079	N/A	.341±.060	N/A
	Extraction Cost (mean±std.)				
MiniBooNE	**42.000±0.000**	45.123±0.091	45.074±0.075	43.095±0.083	50.000±0.000
Forest	38.040±0.709	49.020±0.655	45.140±0.959	42.094±0.751	**34.523±0.751**
Cifar	250.000±0.000	239.950±0.679	**225.293±0.506**	228.853±0.000	250.000±0.000
NUS	**1.000±.000**	1.000±.000	N/A	1.000±.000	N/A
Scene	0.829±.052	0.987±.065	N/A	**0.715±.033**	N/A

are given by attributes 17,18,19 and 20, which can be denoted as 4 different modalities. And the objective is to correctly classify the soil type. The cost of each modality sets equal in raw data.

The small datasets are with partitioned modalities, thus we compare DMP with the novel used modal extraction methods, i.e., LP and DAG used in (Wang et al., 2015). It is notable that different modal costs are same in the 2 small datasets, thus, the prediction cost budget is set as $\{1, 2, 3\}$ in letter and $\{1, 2, 3, 4\}$ in Landsat for simplicity, and we randomly select a modality for the "initial modality". Table 1 records the minimum error rates and corresponding average extraction cost of DMP and compared methods. From Table 1, it clearly reveals that on 2 small real world datasets, the average error rates of DMP are the least, though the average used modalities are not least in Landsat dataset, we can find that the average error rate of DMP is lower with the same average used modalities condition from Fig. 4, which means DMP can achieve the best performance with least modal extraction cost. To investigate the performance of compared modal extraction methods when modal extraction cost changes,we conduct additional experiments and record more results, and a detailed classification performance on different average used modalities are recorded in Fig. 4. From these subplots in Fig. 4, we can find that the error rates of DMP decrease faster than the compared methods, i.e., in Letter dataset, when the cost budget is 2, the downtrend of DMP is significant, which means that the DMP selects the most discriminative modalities based on the previous selected modalities with least cost. Besides, DMP always achieved the best classification performance with the same overall cost at last.

4.2. Large Datasets

Then, we examine the performance of the DMP using 5 higher dimensional sets of data previously used to compare budgeted learning performance.

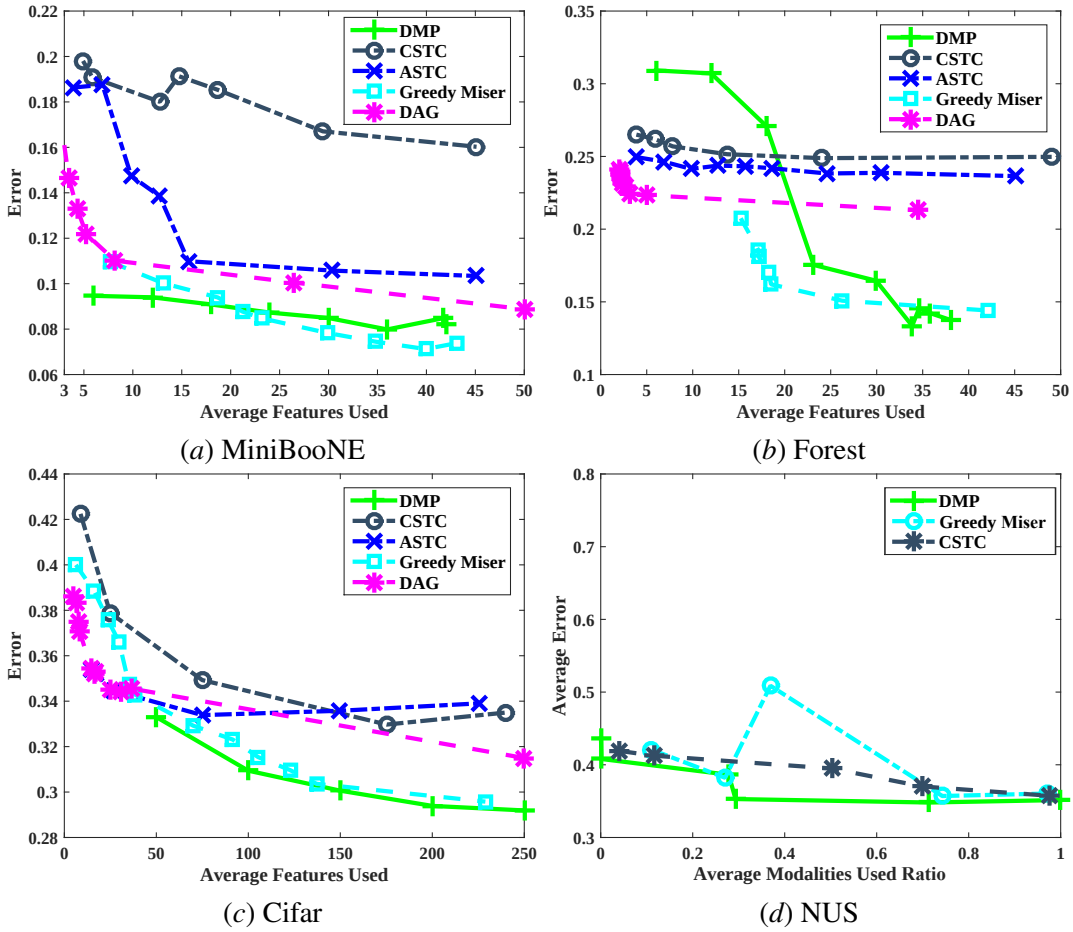


Figure 5: Average number of modalities acquired vs. test error comparison.

3 large datasets are used in used in (Wang et al., 2015): MiniBooNE (Asuncion and Newman, 2007) is a binary classification task, with the goal of distinguishing electron neutrinos (signal) from muon neutrinos (background). Each data point consists of 50 experimental particle identification variables (features). Forest (Asuncion and Newman, 2007) contains cartographic variables to predict 7 forest cover types. Each example contains 54 (10 continuous and 44 binary) features. Cifar-10 (Krizhevsky and Hinton, 2009) consists of 32x32 color images in 10 classes. 400 features for each image are extracted using technique described in (Coates and Ng, 2011). However, there have no partitioned modalities in these cases, and the dimensionality of the data (between 50 and 400 features) makes exhaustive subset construction computationally infeasible, thus, we greedily construct modal subsets according to the calculated information gain of each features as in (Wang et al., 2015), then train the DMP over the partitioned modal features. Note that there have no explicit feature costs, instead, we use the length of the modal features as the cost setting.

Besides, another two large datasets with specific modal partition and extraction cost are also compared: i.e., Scene Li et al. (2001) and NUS Chua et al. (2009). Scene contains $1,000$ images of ten categories, each category containing 100 images and there are 60 features in 28 groups are

generated for each image, including 12 RGB colors, 9 features describing mean hue of 3×3 sub-images, 16 histogram features, 1 contrast features, 4 directionality features, 2 coarseness features, 12 Gabor features and 4 Haralick textures. Features in the same group are extracted together. Feature extraction time cost varies from 4.0×10^{-4} second (RGB color) to 9.6×10^{-2} second (Gabor features). It is notable that the feature extraction time cost of all features is over 1.5 seconds, which is larger than the prediction time of SVM, i.e., around 0.5 second. This indicates that the feature extraction time cost dominates the overall costs during testing phase. Similarly, we have another large scale image datasets compared in the empirical studies: NUS subset contains 19,518 images of ten categories, and six groups of features extracted from these images, including 64 color histogram features, 144 dimension color correlogram, 73 edge direction histogram features, 128 wavelet texture features, 225 block-wise color moments and 500 bag of words based on SIFT descriptions. For DMP can also compare with feature extraction methods, thus we add the novel feature extraction methods, i.e., CSTC, ASTC, Greedy Miser. And record the minimum error rates and corresponding average extraction cost of DMP and compared methods in Table 1. From the results, it clearly reveals that on 4 large real world datasets, i.e., Forest, Cifar, NUS, Scene, the average error rates of DMP are the least and for the rest datasets, we achieves the runner-up, considering that DMP is a modal extraction method and extract a feature subset one step, which is more consistent with real application, e.g., medical diagnosis, while CSTC, ASTC, Greedy Miser are feature extraction methods, and need acquire the modalities which include the extracted features indeed. However, the extraction costs of DMP are also competitive in 5 large datasets, which means more less extraction cost in real. Meanwhile, to investigate the performance of compared modal/feature extraction methods when extraction cost changes, we conduct additional experiments changing the extraction cost and record the classification results on different average used modalities/features in 5. Due to the page limitation, we list 4 datasets here, i.e., MiniBooNE, Forest, Cifar, NUS. From these subplots in Fig. 5, we have the similar results with small dataset, that the error rates of DMP decrease faster than the compared methods, i.e., Forest, Cifar datasets, which means the DMP can expand to large datasets well.

5. Conclusion

Development of data collection ability has spawned the multi-modal learning, while the cost of feature extraction in multi-modalities becomes the major burden in testing with multi-modal learners. In this paper, we propose the instance specific Discriminative Modal Pursuit (DMP) approach, which can automatically choose the most discriminative feature group with one single modality separately for each concerned test instances by steps. As a consequence, the overall cost of modal feature extraction in the test phase can be significantly reduced. Instead of jointly optimizing a highly non-convex empirical risk minimization problem, we turn this problem to an end-to-end learner, which can give the label predictions as well as predict which modality to extract simultaneously. In this way, our DMP approach can reduce the extraction cost, works with limited feature value acquisition budget, and moreover, it can balance the classification performance and modal extraction cost by extracting different modalities for different unseen instances. Experiments on 7 real-world datasets validate the effectiveness of our methods compared with other state-of-the-art methods. It is interesting to place the multi-modal feature extraction problem into reinforcement learning environment, and how to transfer the learned model between different domain is also an interesting future work.

References

Arthur Asuncion and David Newman. Uci machine learning repository. Technical report, University of California, Irvine, 2007.

Boris Babenko, Ming-Hsuan Yang, and Serge Belongie. Robust object tracking with online multiple instance learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(8):1619–1632, 2011.

Minmin Chen, Zhixiang Xu, Kilian Q. Weinberger, Olivier Chapelle, and Dor Kedem. Classifier cascade: Tradeoff between accuracy and feature evaluation cost. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pages 235–242, La Palma, Canary Islands, 2012.

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 the ACM International Conference on Image and Video Retrieval*, page Article No.48, Santorini, Greece, 2009.

Adam Coates and Andrew Y Ng. The importance of encoding versus training with sparse coding and vector quantization. In *Proceedings of the 28th International Conference on Machine Learning*, pages 921–928, Bellevue, Washington, 2011.

Tianshi Gao and Daphne Koller. Active classification based on value of classifier. In *Advances in Neural Information Processing Systems 24*, pages 1062–1070, Granada, Spain, 2011.

Claudio Gentile, Shuai Li, and Giovanni Zappella. Online clustering of bandits. In *Proceedings of the 31st International Conference on Machine Learning*, pages 757–765, Beijing, China, 2014.

Alex Graves and Jurgen Schmidhuber. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 18(5):602–610, 2005.

Ling Jian, Jundong Li1, Kai Shu, and Huan Liu. Multi-label informed feature selection. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, pages 1627–1633, New York, NY, 2016.

Kirill Trapeznikov Joseph Wang, Tolga Bolukbasi and Venkatesh Saligrama. Model selection by linear programming. In *Proceedings of the 12nd European Conference on Computer Vision*, pages 647–662, Zurich, Switzerland, 2014.

Pallika Kanani and Prem Melville. Prediction-time active feature-value acquisition for cost-effective customer targeting. *Proceedings of the Workshop on Cost Sensitive Learning Neural Information Processing Systems 22*, 2008.

Sergey Karayev, Mario J Fritz, and Trevor Darrell. Dynamic feature selection for classification on a budget. In *Proceedings of the 30th International Conference on Machine Learning: Workshop on Prediction with Sequential Models*, Atlanta, Georgia, 2013.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, Department of Computer Science, University of Toronto, 2009.

Matt J Kusner, Wenlin Chen, Quan Zhou, Zhixiang Eddie Xu, Kilian Q Weinberger, and Yixin Chen. Feature-cost sensitive learning with submodular trees of classifiers. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence*, pages 1939–1945, Quebec City, Canada, 2014.

Jia Li, James Ze Wang, and Gio Wiederhold. SIMPLIcity: Semantics-sensitive integrated matching for picture libraries. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(9): 947–963, 2001.

Li-Ping Liu, Yu Yang, Yuan Jiang, and Zhi hua Zhou. TEFE: A time-efficient approach to feature extraction. In *Proceedings of the 8th International Conference on Data Mining*, pages 423–432, Pisa, Italy, 2008.

Feng Nan, Joseph Wang, and Venkatesh Saligrama. Feature-budgeted random forest. pages 1983–1991, Lille, France, 2015.

Alessandro Panella and Piotr Gmytrasiewicz. Bayesian learning of other agents' finite controllers for interactive pomdps. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence*, pages 2530–2536, Phoenix, AZ, 2016.

Xiaonan Song and Haiping Lu. Multilinear regression for embedded feature selection with application to fmri analysis. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, pages 2562–2568, San Francisco, California, 2017.

Paul Viola and Michael Jones. Robust real-time object detection. *International Journal of Computer Vision*, 2(57):34–47, 2001.

Joseph Wang, Kirill Trapeznikov, and Venkatesh Saligrama. An lp for sequential learning under budgets. In *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics*, pages 987–995, Reykjavik, Iceland, 2014.

Joseph Wang, Kirill Trapeznikov, and Venkatesh Saligrama. Efficient learning by directed acyclic graph for resource constrained prediction. In *Advances in Neural Information Processing Systems 28*, pages 2152–2160, Montreal, Canada, 2015.

Zhixiang Xu, Kilian Weinberger, and Olivier Chapelle. The greedy miser: Learning under test-time budgets. In *Proceedings of the 29th International Conference on Machine Learning*, pages 987–995, Edinburgh, UK, 2012.

Zhixiang Eddie Xu, Matt J Kusner, Kilian Q Weinberger, and Minmin Chen. Cost-sensitive tree of classifiers. In *Proceedings of the 30th International Conference on Machine Learning*, pages 133–141, Atlanta, Georgia, 2013.

Lijun Zhang, Tianbao Yang, Rong Jin, Yichi Xiao, and Zhi hua Zhou. Online stochastic linear optimization under one-bit feedback. In *Proceedings of the 33rd International Conference on Machine Learning*, pages 392–401, New York, NY, 2016.

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.