Discover Multiple Novel Labels in Multi-Instance Multi-Label Learning

Shreya Mandavilli
Kartikey Pant
Aiswarya Sunil
Srujana Peddinti

Abstract

Multi-instance multi-label learning (MIML) is a learning paradigm where an object is represented by a bag of instances and each bag is associated with multiple labels.

In this project, we are implementing the first approach to discover multiple novel labels in MIML problem using an efficient Augmented Lagrangian Optimization, which has a bag-dependent loss term and a bag-independent clustering regularization term, enabling the known labels and multiple novel labels to be modeled simultaneously.

Introduction

Traditionally : In supervised learning, an object is a single instance associated with a single label.

Single-Instance Single-Label -> SISL !

Realistically : An object can be simultaneously associated with multiple labels, whereas the object can be described by multiple instances.

Multiple-Instance Multiple-Label -> MIML !

Ideally, a learned model should be able to detect novel instances and thus classify these instances into one of the multiple novel labels.
Example application of such a feature : distinguish multiple new voices in bird song recognition, *corresponding to different birds which are the new arrivals*

Introduction

Why DMNL ?

1. It discovers multiple novel labels, rather than dealing with one novel label only.(in contrast to Pham et. al.)
2. Its computational cost increases linearly, rather than increasing exponentially w.r.t. the number of bag labels.(due to discriminative nature in contrast to the probabilistic nature of other approaches)

Preliminaries

Symbols and their Meanings	
Symbols	**Meanings**
χ	Instance Feature Space
$\mathrm{L} = \{l_1, ..., l_c\}$	target label set of size c
$\overline{L} = \{l_{c+1}, ...l_{c+k}\}$	k novel labels
$\hat{L} = L \cup \overline{L}$	combined set of labels
$\mathrm{D} = \{(X_1, y_1), ..., (X_m, y_m)\}$ $X_i = \{x_{i,1}; .....; x_{i,z_i}\}$ $y_i = [y_{i,1}, ..., y_{i,c}] \in \{0, 1\}^c$	training set of size m bag of z_i instances each $x_{i,j} \in \chi$ observed bag label vector. If bag i belongs to l_j, then $y_{i,j} = 1$; otherwise $y_{i,j} = 0$.
$\hat{y}_{i,j} = [\hat{y}_{i,j,1}, .., \hat{y}_{i,j,c+k}] \in \{0,1\}^{c+k}$	unknown instance label vector for instance j in bag i.
$\mathrm{A} = [X_1; ...; X_m]$	all-instance matrix which is the concatenation of instances from all bags
$\hat{Y}_i = [\hat{y}_{i,1}; ...; \hat{y}_{i,z_i}]$ $\widetilde{Y} = [\hat{Y}_1; ...; \hat{Y}_m]$	instance label matrix of bag *i*. concatenation of all instance label vectors

We follow a common assumption in the MIML setting, i.e., each instance belongs to a single label only. Hence we have,

$$\sum_{l=1}^{c+k} \hat{y}_{i,j,l} = 1$$

Problem Formalization

Problem Definition : Given a training set D, consists of bags with known labels, the problem of discovering multiple novel labels in MIML learning is to detect previously unknown labels (i.e., novel labels) in each bag in the training set, and build a model that can predict bag labels from the set of known and novel labels for a previously unseen bag.

Two levels of tackling the problem :

1. **Instance-level annotation** : It is a mapping from an instance to a label (in the set of known labels and novel labels) f : $\chi \rightarrow \hat{L}$
2. **Bag-level prediction task** : It is a mapping from a bag to a set of labels $\Psi : 2^{\chi} \rightarrow 2^{\hat{L}}$.

Assumptions : Labels of a bag includes all instance labels in that bag, i.e, if any one instance has the label, we assign that label for the entire bag also.

Discover Multiple Novel Labels : Proposition 1

Proposition 1 : $y_i = \beta_i^T \hat{Y}_i$ where $\beta_i = [\beta_{i,1}; ...; \beta_{i,z_i}]$, with each $\beta_{i,j} = 1/\Sigma_{l=1}^{c}(\mathbb{I}(\hat{y}_{i,j,l} = 1)\Sigma_{q=1}^{z_i}\hat{y}_{i,q,l})$

- In Proposition 1, each instance label contributes to the bag label, and β_i corresponds to the contribution weights and $\beta_{i,j}$ is the weight of the j-th instance in the i-th bag.
- According to a rescaling (Zhou and Liu 2010) strategy, suppose there are n_l instances with the l-th label in the bag, then the weight for each of them will be $1/n_l$.
- This proposition also satisfies $y_i = \vee_{j=1}^{z_i}\hat{y}_{i,j}$
- **Cluster structure assumption** : There exists a prototype for each label, and instances with the same label are close to the label prototype, and far away from the prototypes of other labels.

Discover Multiple Novel Labels : Proposition 2

Proposition 2 : Given prototype p_l of label l, instance label matrix $\tilde{Y}$ is solution to

$$\min_{G} \Sigma_{i=1}^{n}\Sigma_{l=1}^{c} G_{i,l}||a_i - p_l||^2 : G \in \{0,1\}^{nxc}, G^T G = S, where, S = diag(1^T G) : \text{Eqn}(1)$$

Since prototype p is unknown, the optimization must avoid calculating p.

Applying the spectral relaxation on Eqn(1), we obtain a relaxed maximization problem. By defining $H = \tilde{Y} S^{-\frac{1}{2}}$: $(1) \Rightarrow \max_{H} trace(H^T A A^T H) : H^T H = I, H \geq 0$

Spectral relaxation of K-means clustering

Given a set of m-dimensional data vectors $a_i, i = 1, ..., n$, we form the m x n data matrix $A = [a_1, ..., a_n]$. A partition π of the date vectors can be written in the following form :

$AE = [A_1, .., A_k], A_i = [a_1^{(i)}, ...a_{s_i}^{(i)}]$

For a given partition π, the associated sum of squares cost function is given by

- $ss(\pi) = \Sigma_{i=1}^{k} \Sigma_{s=1}^{s_i} ||a_s^{(i)} - m_i||^2, m_i = \Sigma_{s=1}^{s_i} a_s^{(i)}/s_i = A_i e/s_i$
- $ss_i \equiv \Sigma_{s=1}^{s_i} ||a_s^{(i)} - m_i||^2 = ||A_i - m_i e^T||_F^2 = ||A_i(I_{s_i} - ee^T/s_i)||_F^2$
  $ss_i = trace(A_i(I_{s_i} - ee^T/s_i)A_i^T) = trace((I_{s_i} - ee^T/s_i)A_i A_i^T)$
- $ss(\pi) = \Sigma_{i=1}^{k} ss_i = \Sigma_{i=1}^{k} (trace(A_i A_i^T) - (\frac{e_T}{\sqrt{(s_i)}}) A_i A_i^T (\frac{e}{\sqrt{(s_i)}}))$
- $LetX = \begin{pmatrix} \frac{e}{\sqrt{(s_1)}} & .. & \\ .. & .. & \\ .. & .. & \frac{e}{\sqrt{(s_k)}} \end{pmatrix}$

The sum of squares cost function can now be written as :

$$ss(\pi) = trace(A^T A) - trace(X_T A_T A X)$$

Approach by combining propositions

Proposition 1 enables us to design a bag-level loss ($\Sigma_{i=1}^{m}||y_i - \beta_i^T \Omega_c(\hat{Y}_i)||^2$), considering the contribution of each instance.

Proposition 2 enables us to design ($-Tr((\tilde{Y}S^{-\frac{1}{2}})^T AA^T \tilde{Y}S^{-\frac{1}{2}}$) as the regularization term considering the cluster structure.

Combining the loss and the regularization term together under the non-negative orthogonal constraints, we obtain the following optimization task :

$$\min_{\tilde{Y}} \Sigma_{i=1}^{m}||y_i - \beta_i^T \Omega_c(\hat{Y}_i)||^2 - \lambda Tr((\tilde{Y}S^{-\frac{1}{2}})^T AA^T \tilde{Y}S^{-\frac{1}{2}} \text{ : Eqn(2)}$$

Optimization

- Substituting $\tilde{Y}$ in Eqn(2) as g(A,W), the final optimization for DMNL is given as follows :
  $\min_{W}$
  $\Sigma_{i=1}^{m} \parallel y_i - \beta_i^T \Omega_c(g(X_i, W)) \parallel^2 - \lambda Tr((g(A, W)S^{-\frac{1}{2}})^T AA^T(g(A, W)S^{-\frac{1}{2}}))$: Eqn(3)

- **Instance-level annotation** : Having learned W, instance $x_{i,j}$ is assigned the label with the maximum predictive value, i.e.,

$$\hat{y}_{i,j,l} = \begin{cases} 1, l = \arg\max_l, g_l, (x_{i,j}, W), l^{'} \in 1, ..., c + k; \\ 0, otherwise \end{cases} : \text{Eqn(4)}$$

Optimization

- Eqn(3), using optimization strategy can be written as :
  $\min_{W,\hat{H}} \phi(W) + \psi\hat{H} : \hat{H}^T\hat{H} = I, \hat{H} = g(A,W)S^{-\frac{1}{2}}$: Eqn(5) ,
  where $\phi(W) = \frac{1}{2}\Sigma_{i=1}^{m} \parallel y_i - \beta_i^T \Omega_c(g(X_i,W)) \parallel^2$ and
  $\psi(\hat{H}) = \frac{1}{2}\Sigma_{i=1}^{m} \parallel y_i - \beta_i^T \Omega_c(H_i S^{\frac{1}{2}}) \parallel^2 - \lambda Tr(\hat{H}^T A A^T \hat{H})$.

- To solve Eqn(5), we use the Lagrangian optimization framework and the augmented lagrangian of Eqn(5) is given by :

  $$\mathcal{L}\ (W, \hat{H}, \wedge) = \phi(W) + \psi(\hat{H}) + \frac{\sigma}{2} \parallel \hat{H} - g(A,W)S^{-\frac{1}{2}} + \wedge \parallel_F^2 + \zeta$$

  where $\parallel . \parallel_F$ is the Frobenius norm, $\wedge$ is the dual form, σ is the penalty parameter and ζ is a constant which can be dropped during the optimization.

Optimization

- Solving Eqn(5), we get :
  $\min_{W,\hat{H},\wedge} \mathcal{L}(\mathrm{W}, \hat{H}, \wedge) : \hat{H}^T\hat{H} = I, \hat{H} \geq 0 : \mathrm{Eqn}(6)$

- Optimizing Eqn(6), w.r.t. W, $\hat{H}$ and $\wedge$ in an alternating manner. Let
  $W^{t+1} = \arg\min_W \mathcal{L}(\mathrm{W}, \hat{H}^t, \wedge^t) ; : \mathrm{Eqn}(7)$
  $\hat{H}^{t+1} = \arg\min_{\hat{H}\geq 0} \mathcal{L}(W^{t+1}, \hat{H}, \wedge^t) : \hat{H}^T\hat{H} = I : \mathrm{Eqn}(8)$
  $\wedge^{t+1} = \wedge^t + \hat{H}^{t+1} - g(A, W^{t+1})S^{-\frac{1}{2}} : \mathrm{Eqn}(9)$

- **Update W.**In orer to obtain Eqn(7), Stochastic Gradient Descent(SGD) is applied. Specifically solving Eqn(7) is equivalent to minimizing :

  $\mathcal{L}_W = \phi(W) + \frac{\sigma}{2} \parallel \hat{H} - g(A, W)S^{-\frac{1}{2}} + \wedge \parallel_F^2.$

Optimization

- **Update** $\hat{H}$. For the optimization task of Eqn(8), there is a non-negative orthogonal constraints $\hat{H}^T\hat{H} = I, \hat{H} \geq 0.\hat{H}^T\hat{H} = I$ is known as the Stiefel manifold, thus we solve Eqn(8) on the Stiefel manifold.We derive the update rule for $\hat{H}$ as :

$$\hat{H}^{t+1} = \hat{H}^t \circ \frac{[\nabla_{\hat{H}}]^+ + \hat{H}([\nabla_{\hat{H}}]^-)^T\hat{H}}{[\nabla_{\hat{H}}]^- + \hat{H}([\nabla_{\hat{H}}]^+)^T\hat{H}} \text{ : Eqn(10)}$$

  where $\nabla_{\hat{H}}$ is the gradient of $\mathcal{L}\ (W^{k+1}, \hat{H}, \wedge^k)$ w.r.t $\hat{H}$; $[\nabla_{\hat{H}}]^+$ and $[\nabla_{\hat{H}}]^-$ satisfy $[\nabla_{\hat{H}}]^+ > 0$, $[\nabla_{\hat{H}}]^- > 0$, $[\nabla_{\hat{H}}] = [\nabla_{\hat{H}}]^+$ - $[\nabla_{\hat{H}}]^-$

- **Stiefel Manifold** : The (compact) Stiefel manifold $V_{n,p}$ p is the set of all p-tuples $(x_1, \ldots, x_p)$ of orthonormal vectors in $\mathbb{R}^n$. If we turn p-tuples into n $\times$ p matrices as follows :

$$(x_1, \ldots, x_p) \mapsto [x_1 \ \cdot \ \cdot \ \cdot \ x_p],$$
$$V_{n,p} = \ \mathrm{X} \in \mathbb{R}^{nxp} : X^T\mathrm{X} = I_p$$

Algorithm 1 DMNL

Input: D, λ, ρ
Output: W
Process:

Initialize W, $\hat{H}$, Λ, S and $\boldsymbol{\beta}_i, i = 1, \cdots, m$;
repeat:
    Update W via SGD; \\Solve Eqn. (7)
    Update $\hat{H}$ via Eqn. (10); \\ Solve Eqn. (8)
    Update Λ via Eqn. (9);
    Predict $\tilde{Y}$ according to Eqn. (4);
    Calculate $\boldsymbol{\beta}_i, i = 1, \cdots, m$, according to Proposition. 1;
    Calculate $S = \mathrm{diag}(\mathbf{1}^\top \tilde{Y})$;
until convergence or the maximum iteration is reached.

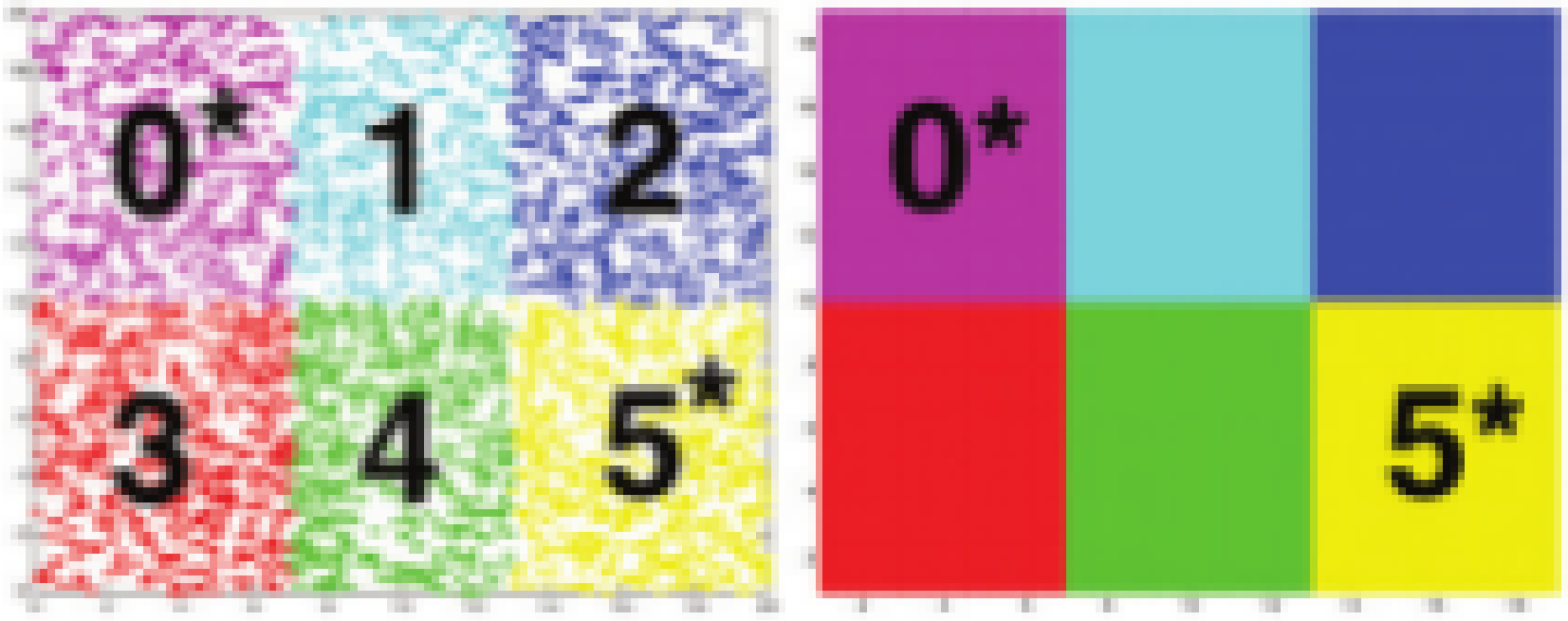
Performance Evaluation

The experiments are done on the following data sets for the evaluation process :

- Toy dataset
- Real datasets
  - MSRCv 2 image dataset(MS)
  - Two letter datasets(Briggs, Fern, and Raich 2012) (i.e., Letter Carroll(LC) and Letter Frost(LF))
  - The MNIST handwritten dataset. (Convert this SISL dataset to MIML dataset)(MN)

Experimental set up :

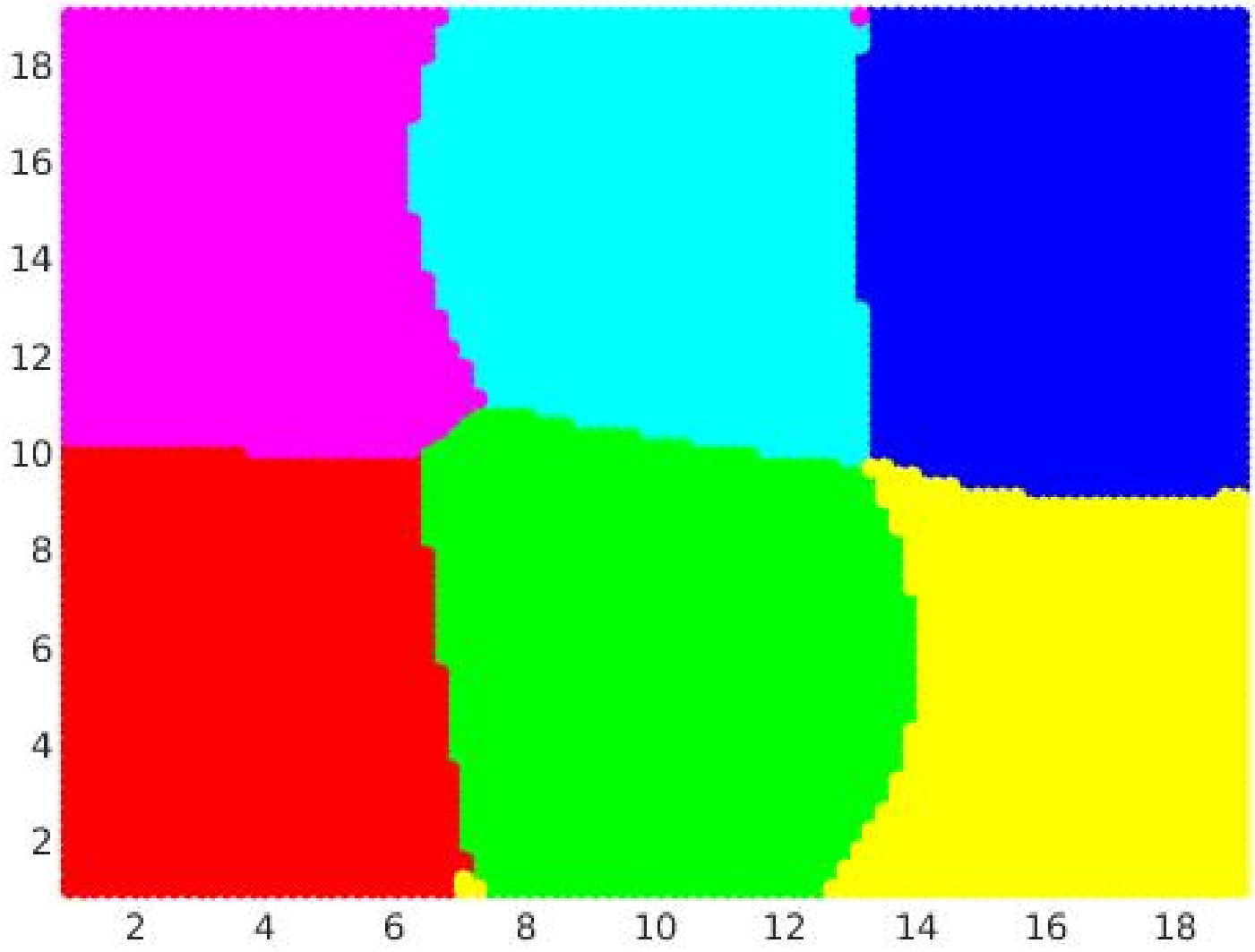
1. If the given dataset is not in MIML format, convert it into MIML format.
   For Eg. MNIST is a single-instance single-label dataset taken from 10 digits. In order to transform it to a MIML format, we randomly sample 200 bags from the 10 digits, resulting each bag having 27.6 instances and 3.09 labels on average.
2. Divide the class labels into know and unknown labels(novel labels) for the taken dataset.
3. Remove the novel labels for training.



(a) Training data (b) Test data

Performance Evaluation

Output



Performance Evaluation : Continued

It indicates that our implementation is significantly similar in performance to the one implemented in the paper.

approach/metric	**Paper Implementation**	**Our Implementation**
Accuracy	0.6670	0.6650
AUC	0.7450	0.7413

Conclusion

We presented the first model for discovering and predicting multiple novel labels in MIML learning. The proposed discriminative model has the following unique feature : the problem is formulated as a non-negative orthogonal con-strained optimization problem that has a bag-dependent loss term and a clustering regularization term which is bag-independent.

This enables both the known labels and the multiple novel labels to be simultaneously modeled. Experiments results validate the effectiveness and the efficiency of our approach on discovering and predicting multiple novel labels in MIML learning.