Multi-Label Learning with Posterior Regularization

Xi Victoria Lin, Sameer Singh, Luheng He, Ben Taskar, Luke Zettlemoyer
Computer Science & Engineering
University of Washington
{xilin,sameer,luheng,lsz}@cs.washington.edu

Abstract

In many multi-label learning problems, especially as the number of labels grow, it is challenging to gather completely annotated data. This work presents a new approach for multi-label learning from incomplete annotations. The main assumption is that because of label correlation, the true label matrix as well as the soft predictions of classifiers shall be approximately low rank. We introduce a posterior regularization technique which enforces soft constraints on the classifiers, regularizing them to prefer sparse and low-rank predictions. Avoiding strict low-rank constraints results in classifiers which better fit the real data. The model can be trained efficiently using EM and stochastic gradient descent. Experiments in both the image and text domains demonstrate the contributions of each modeling assumption and show that the proposed approach achieves state-of-the-art performance on a number of challenging datasets.

1 Introduction

In multi-label classification, the goal is to assign a set of labels to each data point. In many such problems, especially as the number of labels grows, it is challenging to gather completely annotated data. For example, consider the commonly studied *delicious* dataset [15] where the goal is to predict which of nearly 1,000 common tags (e.g., "wiki" or "history") can be assigned to a web page. Although delicious users tag sites regularly, providing positive labels, no effort is made to ensure completeness. Such missing labels are common in many domains, ranging from image annotation to protein function prediction, and provide a challenge for learning.

The missing-label problem has been widely known among the community. Previous work have addressed it the following ways. (1) *Cost-sensitive learning* assigns higher penalties to false negatives than false positives [4]. In particular, SVM-VT [16] extends this idea by adding a tolerance parameter in the hinge loss to errors caused by missing synonyms and hypernyms. (2) *Learning to rank* encourages classifiers to rank the observed labels higher than the rest [18, 4]. The implicit assumption is that labels tagged by users are more important, while the relative importance of the unobserved labels is uninformed, on which the classifiers are not forced to make a decision. (3) *Leveraging label correlation* let the decisions over different labels interfere by either regularizing the model parameters, e.g., applying the low-rank constraint [12, 20] or explicitly modeling pairwise correlations [9, 2]. (4) *Instance-based learning* propagate labels from annotated instances to the new ones based on similarities in the feature/label space or both [8, 19].

Our approach starts from modeling the patterns that the ideal labelings of such problems should have. First, we expect the assigned labels to be *sparse*, since on average each instance is associated with a small subset of the possible labels. Second, labels are correlated and groups of labels tend to co-occur, for example "train" and "railroad" are likely to appear together in images. Hence we also expect a complete label matrix to be *approximately* low-rank. Such correlations might be concealed if all the missing labels are treated as negative annotations. Our objective is to train an inductive classifier using both the provided labels and the structural assumptions. Following recent work on posterior regularization [6], we formulate the *sparse* and *low-rank* assumptions as soft constraints on the classification posterior, regularizing it to be close to a low-rank and sparse factorization of

the observed label matrix. Our method for achieving this goal can also be seen as jointly optimizing model training and label completion, with the two components mutually benefit each other.

Compared to previous work which introduces strict low-rank parameters [20], our model performs superior in terms of $F1$-measure. We believe this comes from the benefits of learning label correlations from the factorized label matrix without changing the original model parameterization. The posterior-regularized model also performs much better than the baseline that does the low-rank and sparse factorization and learning in a stage-wise manner, confirming previous research that label matrix factorization can benefit from instance features as side information [19]. It also shows that although the label matrix factorization acts as an effective posterior regularizer, it cannot substitute the gold labels, i.e. the beliefs of the learners w.r.t. the gold labels shall not be softened.

2 Model

We assume access to a training data set with N examples $\{(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_N, \mathbf{y}_N)\}$. For a problem with L possible labels, $y_{ij} = 1$ if data point i is tagged with label j; $y_{ij} = 0$ otherwise. We represent the example features and labels as matrices X and Y, of dimensions $N \times D$ and $N \times L$, respectively. Let $\mathcal{O} \subseteq \{1 \ldots N\} \times \{1 \ldots L\}$ be the index set of observed labels in Y. We also assume that Y contains missing labels, such that some of the 0's in Y have high probability of being 1.

2.1 Probabilistic Classifier

We will learn a classifier $P_\theta : R^D \rightarrow (0,1)^L$, where $P_\theta(\mathbf{x}_i, j)$ is the predicted probability that $\mathbf{y}_i^j$ is 1. Here we use the binary relevance (BR) model with logistic regression [14] to define P_θ. The BR model uses independent logistic regression classifiers for each label in $\{1 \ldots L\}$. The resulting parameters θ can be represented as matrix of size $D \times L$ and $P_\theta(\mathbf{x}) = \text{sigm}(\mathbf{x} \times \theta)$, where $\text{sigm}(v)$ is the element-wise sigmoid function ($\frac{1}{1+e^{-v}}$). The likelihood of the model factors over labels:

$$\mathbb{L}(Y_i, P_\theta(\mathbf{x}_i)) = \sum_{j=1}^{L} \mathcal{L}(Y_{ij}, P_\theta(\mathbf{x}_i, j)), \quad \text{where } \mathcal{L}(y, p) = -\log\left(p^y (1-p)^{1-y}\right) \tag{1}$$

If Y is only partially observed, as marked by an index set $\mathcal{O}$, we can rewrite the objective as below, where $\mathcal{O}$ consists of all the labels provided by the annotators.

$$\min_\theta \sum_{(i,j) \in \mathcal{O}} \mathcal{L}(Y_{ij}, P_\theta(\mathbf{x}_i, j)) + \lambda_P ||\theta||_2^2 \tag{2}$$

2.2 Sparse and Low-rank Factorization

The missing labels to deal with are one-sided: all positive annotations $Y_{ij} = 1$ can be trusted, but $Y_{ij} = 0$ should not be directly used as negative annotations. Hence to predict the true label structure of an item, we need to identify which of the missing cells are false negatives due to annotation noise.

Let Q be the belief matrix that captures the factorization of the labels, which is $N \times L$ in size, and each $Q_{ij} \in [0,1]$ is the probability that item i gets label j. Based on the structural assumptions introduced in § 1, Q is a sparse and low-rank matrix close to Y. The objective for computing Q is:

$$\min_{U,V} \sum_{(i,j) \in \mathcal{O}} \mathcal{L}(Y_{ij}, Q_{ij}) + \lambda_Q |Q| + \lambda_U ||U||_2^2 + \lambda_V ||V||_2^2 \tag{3}$$

where $Q_{ij} = \text{sigm}(U_i \cdot V_j)$ and U, V are $N \times K$ and $L \times K$ sized matrices ($K << L, N$). $\mathcal{L}$ is the negative log-likelihood as defined in Eq. 1. $\lambda_Q |Q|$ is the L_1 regularization that encourages sparsity.

2.3 Joint Inference via Posterior Regularization

To allow the structural assumptions about the label matrix to influence supervision, and to leverage feature information for label completion, we need to enforce consistency between P_θ and Q. We introduce the following objective that minimizes the KL divergence between them, which is defined

Table 1: Example-averaged and label-averaged F1-measures of the methods on full annotations.

	bibex		**delicious**		**bookmarks**		**eurlex**		**iaprtc12**		**espgame**	
	Exp	Lab	Exp	Lab	Exp	Lab	Exp	Lab	Exp	Lab	Exp	Lab
BR	37.2	26.8	26.5	10.2	30.7	21.9	43.0	28.8	44.1	20.9	28.2	20.8
LPOS	43.9	37.0	29.0	**12.0**	34.6	**24.3**	33.3	27.7	40.4	21.6	27.9	21.1
TC	37.7	30.0	27.8	5.6	30.3	20.2	39.4	26.7	**45.3**	21.4	**29.6**	21.1
LEML	39.0	28.3	**35.3**	9.7	31.0	13.1	**43.9**	18.9	44.5	15.8	25.5	9.6
PRLR	**44.2**	**37.2**	33.3	**12.0**	**34.9**	23.0	40.2	**29.7**	40.1	**27.2**	28.6	**22.2**

as the sum of the element-wise Bernoulli divergences, i.e. $\mathrm{KL}(Q \parallel P_\theta) = \sum_{i=1}^{N} \sum_{j=1}^{L} \mathrm{KL}(Q_{ij} \parallel P_\theta(\mathbf{x}_i, j))$. The joint objective is then:

$$\min_{U,V,\theta} \quad \mathrm{KL}(Q \parallel P_\theta) + \beta_P \sum_{(i,j)\in\mathcal{O}} \mathcal{L}(Y_{ij}, P_\theta(\mathbf{x}_i, j)) + \lambda_P |P_\theta| + \lambda_\theta ||\theta||_2^2 + \sum_{(i,j)\in\mathcal{O}} \mathcal{L}(Y_{ij}, Q_{ij}) + \lambda_Q |Q| + \lambda_U ||U||_2^2 + \lambda_V ||V||_2^2. \tag{4}$$

Here β_P, λ_P, λ_Q are the problem-specific hyperparameters and λ_θ, λ_U, λ_V are regularization constants. β_P specifies the tradeoff between learning from the encoding distribution (Q) and the observed labels ($Y_{ij}, (i,j) \in \mathcal{O}$). λ_P, λ_Q control the sparsity of P_θ and Q, respectively. An illustration of the joint model is shown in the supplementary material.

3 Experiments

Datasets We use six benchmark multi-label learning datasets[1] with medium to large number of labels. Four of the datasets are from the text labeling domain (*bibtex*, *delicious*, *bookmarks*, *eurlex*) and the rest two are from the image annotation domain (*corel5k*, *espgame*). To further measure the influence of learning with partial labels, we also randomly partitioned the positive labels of each dataset into five subsets, and train classifiers with 20%, 40%, 60%, and 80% of the given labels. The classifiers trained this way are evaluated against all labels.

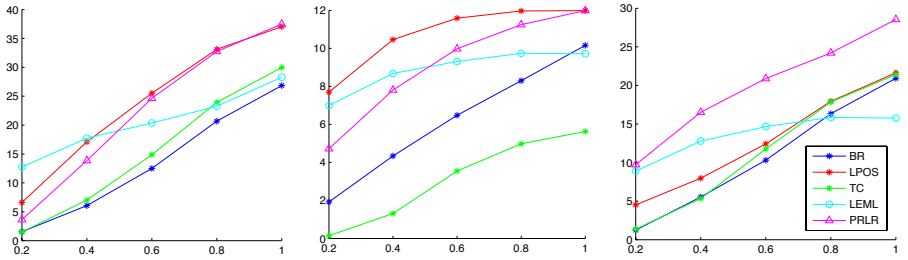
We compare our method (PRLR, for *Posterior-Regularized Low-Rank*) with four baselines. BR, LPOS and TC are proposed to measure the performance of the classifier with certain modeling assumptions removed. LEML is a state-of-the-art large-scale multi-label learning algorithm leveraging label correlations. We tune all the hyper-parameters based on a held-out validation dataset.

- BR, the naive *binary relevance* model [14], consists of an independent logistic regression classifier for each label, using all examples with that label as positive examples and the rest as negative.
- LPOS, or *Learning from Positives Only with Sparsity*, trains a logistic regression classifier for each label independently, with the loss only defined over the observed labels (similar to [1, 3, 4]). In addition we include an $L1$-norm regularization on the posterior to enforce sparsity.
- TC, or *Tag Completion*, first factorizes the label matrix via the optimization program described in Eq. 3, followed by training a BR model with the completed label matrix Q.
- LEML, or *Low rank Empirical risk minimization for Multi-Label Learning* [20] learns a multi-label classifier parameterized by θ that minimizes the empirical loss under the constraint that θ is low-rank. We reimplemented their algorithm ourselves. and obtains top-k accuracy within 1% of those reported in the original paper.

Results over complete annotations Table 1 presents the evaluation results of PRLR in comparison to the four baselines (BR, LPOS, TC and LEML) for the full labels setting. PRLR consistently performs well on all the datasets. Moreover, it has the effect of boosting recall and label coverage significantly while maintaining good precision.[2]

[1]The text datasets are available on `http://mulan.sourceforge.net` and the image datasets on `http://lear.inrialpes.fr/data`.

[2]Since the benchmark datasets may also suffer from missing labels, we are under-estimating precision and over-estimating recall. Nevertheless, the relative difference between the models are still informative.



(a) Bibtex (b) Delicious (c) Corel5k

Figure 1: Comparison of PRLR (magenta curve) with baselines over partial annotations on *bibtex*, *delicious* and *corel5k*.

Among the three baselines, BR in general achieves high precision but suffers a much lower recall, confirming that the naive classifiers are sensitive to the label bias. LPOS performs better than BR on text but worse on the image datasets. On the other hand, learning from the completed labels (TC) marginally improves or is worse than BR on text, but performs well for images.

Interestingly, it appears that the performances of LPOS and TC are complementary to each other. For datasets where learning with positives only does poorly, the annotations provided are not descriptive enough hence label completion is necessary; on the other hand, the fact that learning with positives only performs strongly indicates that the annotations are indeed complete and adding the TC component may generate false positives that hurts the training. PRLR, on the other hand, is able to take advantage of both approaches.

The variation of the performances of LPOS and TC on the text datasets vs. the image datasets seems to indicate that the text datasets are in general more "complete" than the image datasets. By manually examing the label space of these datasets we found that the labels of the text datasets contains more "categorical" words ("branding", "coding", "communication","creativity", etc.), while a large proportion of the labels for the image datasets correspond to physical objects and their properties ("dog", "flower", "girl", "red", "branch"), which are more likely to be ignored.

LEML achieves the best example averaged F1-measure on *delicious* and *eurlex*, the datasets that have the largest number of possible labels among the ones we choose to test on. In such problems, as the label space grows, the correlation between labels becomes more significant; thus the parameter matrix of a linear prediction model is likely to have low-rank structure, and exact factorization methods will perform well. However, it in general suffers from low label-averaged metrics.

Results over partial annotations Figure 1 shows the change of label-averaged F1-measures as the proportion of revealed labels, denoted by η, varies among 20%, 40%, 60%, 80%, 100%. The performances of all methods increase as more labels are revealed. The slope of the LEML curve is much different from the rest since we are taking top-k predictions of this method while using a 0.5 score threshold for the others. We observed that PRLR and LPOS in general performs better than BR and TC as η changes. The performance of LPOS and PRLR is very close on *bibtex*, with LPOS slightly better when very small proportion of the labels are revealed. On *delicious*, the difference is more clear when $\eta < 1$ On *corel5k*, the image annotation dataset, PRLR outperforms the other method by a large margin.

4 Conclusions

Most large-scale multi-label learning problems today have label space that spans a large number of natural language words. Generating complete annotated data for these problems are difficult. We presented a new approach for multi-label learning from incomplete annotations. By incorporating soft posterior constraints, we encourage probabilistic predictions that are low-rank and sparse (after thresholding), which results in more effective classification models. Experiments in the image and text domains demonstrated the contributions of each modeling assumption. Our learning approach was general, in that it is possible to change the family of prediction models and the posterior constraints to allow more sophisticated modeling. It would be interesting to study the effect of posterior regularization for other types of ML models, such as neural networks.

References

[1] Agrawal, R., Gupta, A., Prabhu, Y., and Varma, M. Multi-label learning with millions of labels: Recommending advertiser bid phrases for web pages. In *Int. Conf. on World Wide Web (WWW)*, 2013.

[2] Bian, Wei, Xie, Bo, and Tao, Dacheng. Corrlog: Correlated logistic models for joint prediction of multiple labels. In *AISTATS*, pp. 109–117, 2012.

[3] Bucak, S. S., Mallapragada, P. K., Jin, R., and Jain, A. K. Efficient multi-label ranking for multi-class learning: Application to object recognition. In *International Conference on Computer Vision (ICCV)*, pp. 2098–2105, 2009.

[4] Bucak, S. S., Jin, R., and Jain, A. K. Multi-label learning with incomplete class assignments. In *CVPR*, pp. 2098–2105, 2011.

[5] Duygulu, P., Barnard, K., de Freitas, J. F. G., and Forsyth, D. A. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *European Conference on Computer Vision*, 2002.

[6] Ganchev, Kuzman, Graça, Joao, Gillenwater, Jennifer, and Taskar, Ben. Posterior regularization for structured latent variable models. *The Journal of Machine Learning Research*, 99:2001–2049, 2010.

[7] Grubinger, Michael, Clough, Paul, Mller, Henning, and Deselaers, Thomas. The IAPR TC-12 benchmark – a new evaluation resource for visual information systems, 2006.

[8] Guillaumin, M., Mensink, T., Verbeek, J., and Schmid, C. Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation matthieu. In *ICCV*, 2009.

[9] Hariharan, B., Zelnik-Manor, L., Vishwanathan, S. V. N., and Varma, M. Large scale max-margin multi-label classification with priors. In *Proceedings of the International Conference on Machine Learning*, June 2010.

[10] Katakis, I., Tsoumakas, G., and Vlahavas, I. Multilabel text classification for automated tag suggestion. In *In: Proceedings of the ECML/PKDD-08 Workshop on Discovery Challenge*, 2008.

[11] Kocev, Dragi, Vens, Celine, Struyf, Jan, and Dzeroski, Saso. Ensembles of multi-objective decision trees. In *ECML*, pp. 624–631, 2007.

[12] Loeff, N. and Farhadi, A. Scene discovery by matrix factorization. In *Proceedings of the 10th European Conference on Computer Vision*, Marseille, France, 2008.

[13] Madjarov, G., Kocev, D., Gjorgjevikj, D., and Deroski, S. An extensive experimental comparison of methods for multi-label learning. *Pattern Recogn.*, 45(9):3084–3104, 2012.

[14] Tsoumakas, G. and Katakis, I. Multi label classification: An overview. *International Journal of Data Warehouse and Mining*, 2007:1–13, 2007.

[15] Tsoumakas, G., Katakis, I., and Vlahavas, I. Effective and efficient multilabel classification in domains with large number of labels. In *ECML/PKDD Workshop on Mining Multidimensional Data*, 2008.

[16] Verma, Yashaswi and Jawahar, C. V. Exploring svm for image annotation inpresence of confusing labels. In *British Machine Vision Conference (BMVC)*, 2013.

[17] von Ahn, Luis and Dabbish, Laura. Labeling images with a computer game. In *CHI*, 2004.

[18] Weston, Jason, Bengio, Samy, and Usunier, Nicolas. Large scale image annotation: Learning to rank with joint word-image embeddings. In *European Conference on Machine Learning*, 2010. URL `http://www.kyb.mpg.de/bs/people/weston/papers/wsabie-ecml.pdf`.

[19] Wu, L., Jin, R., and Jain, A. K. Tag completion for image retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(3):716–727, 2013.

[20] Yu, H., Jain, P., Kar, P., and Dhillon, I. Large-scale multi-label learning with missing labels. In *Proceedings of the International Conference on Machine Learning*, Beijing, China, 2014.

Table 2: Dataset Statistics: We show the number of training (N) and test ($\#test$) items, number of features (D), number of labels (L) and the average number of labels per example l_e.

Dataset	N	$\#test$	D	L	l_e
bibtex	4,880	2,515	1,836	159	2.40
delicious	12,920	3,185	500	983	19.02
bookmarks	60,000	27,856	2,150	208	2.03
eurlex	17,413	1,935	5,000	3,993	5.32
corel5k	4,500	499	37,152	260	3.52
espgame	18,689	2,081	37,152	268	4.70

A Dataset Description

The datasets include:

bibtex [10] contains metadata for bibtex items, such as the title of the paper, the authors, venue, etc. The tags are provided by users through the Bibsonomy system (`http://www.bibsonomy.org`).

bookmarks [10] contains metadata for bookmark items like the URL of the web page. The tags are provided by users through Bibsonomy.

delicious [15] contains the text content of web pages. The tags are acquired from Delicious (`https://delicious.com`), a social bookmarking website.

corel5K [5] contains images from the larger Corel CD set. Each image is annotated with an average of 3.5 keywords, and the keywords that appear in both the train and the test set are used as tags.

iaprtc-12 [7] contains images of natural scenes that include cities, landscapes, people, animals, different sports and actions, and many other aspects of contemporary life. The tags are compiled from free-flowing text captions.

espgame [17] contains a variety of images collected from the ESP collaborative image tagging task. During the game, labels assigned to an image by two players without communicating are accepted as its tags, resulting in a challenging dataset for automatic labeling.

A summary of the statistics of the datasets is provided in table 2.

B Posterior Regularization Low-rank

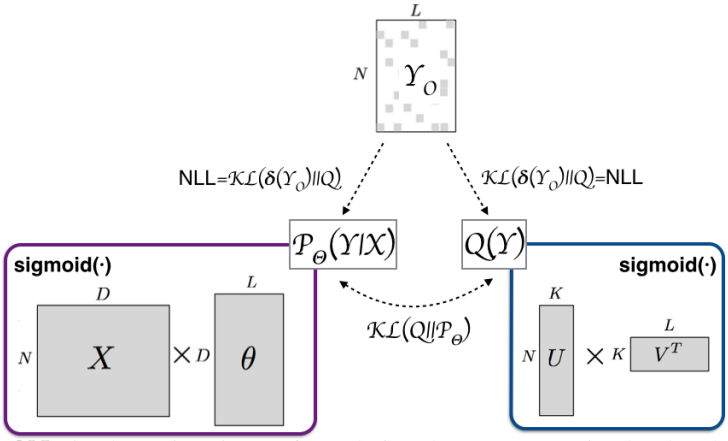


Figure 2: Three NL-sized matrices in our formulation that are constrained to be close: (a) a sparse set of annotations Y such that only the provided labels are used for training, (b) a low-rank, sparse completion of the labels Q represented by rank-K matrices U and V, and (c) posteriors of an inductive probabilistic model P_θ that uses parameter matrix θ to predict the labels given features X.