Evaluating Multi-label Classifiers with Noisy Labels

Anonymous Authors[1]

1. Supplementary Material

1.1. Dataset Statistics

We use five well-studied MLC datasets. We show the dataset statistics in Table 1: the five datasets vary in the input type, number of samples, number of labels, and number of input features; additionally, some datasets have more imbalanced classes than the others.

1.2. Evaluation Metrics

We use the following metrics in the experimental evaluation.

F1-scores. We denote the number of true positives by tp, the number of false positives by fp, and the number of false negatives by fn. An F1-score is defined as follows:

$$F_1 = \frac{2tp}{2tp + (fp + fn)}$$

The example-based F1-score calculates F1-scores for all test samples individually and take the average over them:

$$\mathrm{eb}F_1 = \frac{1}{N} \sum_{i=1}^{N} \frac{\sum_{j=1}^{L} 2y_j^i \hat{y}_j^i}{\sum_{j=1}^{L} y_j^i + \sum_{j=1}^{L} \hat{y}_j^i}$$

where N is the number of test samples, y_j^i is the j-th ground-truth label of test sample i and $\hat{y}_j^i$ is the j-th predicted label of test sample i.

The micro-averaged F1-score sums up individual true positives, false positives, and false negatives of all predication outcomes, and use them to compute an F1-score:

$$\mathrm{mi}F_1 = \frac{\sum_{l=1}^{L} \sum_{i=1}^{N} 2y_l^i \hat{y}_l^i}{\sum_{l=1}^{L} \sum_{i=1}^{N} [2y_l^i \hat{y}_l^i + (1 - y_l^i)\hat{y}_l^i + y_l^i(1 - \hat{y}_l^i)]}$$

The macro-averaged F1-score is the averaged F1-score over all label classes:

$$\mathrm{ma}F_1 = \frac{1}{L} \sum_{l=1}^{L} \frac{\sum_{i=1}^{N} 2y_l^i \hat{y}_l^i}{\sum_{i=1}^{N} [2y_l^i \hat{y}_l^i + (1 - y_l^i)\hat{y}_l^i + y_l^i(1 - \hat{y}_l^i)]}$$

1.3. Additional Experiments

1.3.1. Visualization

In Figure 1, we present a t-SNE visualization (Van der Maaten & Hinton, 2008) of label embeddings of *Pascal VOC* obtained from various settings. With enough training data, starting to learn with randomly initialized label embeddings already produces reasonable results. However, with word embeddings, we are able to obtain more precise label clusterings. The word embedding clusterings (middle) and the regularized word embedding clusterings (right) look similar. That said, regularized word embeddings clusterings are more accurate in the context of *Pascal VOC*; they begin from the word embeddings, but become more relevant to the dataset through the training process. Since the objects falling into the same category have more co-occurrences, they are closer to each other in the embedding space.

1.3.2. When CbMLC fails

We briefly discuss when context-based regularization fails. CbMLC is highly dependent on the quality of pretrained word embeddings. We run it on an additional dataset, *ebird* (Sullivan et al., 2009), which is a bird distribution dataset across the world. Given a set of environmental features $\boldsymbol{x}$ associated to a location, the task is to predict what bird species $\boldsymbol{y}$ there are at the given location. Again, we retrieve species names from GloVe. We summarize the results in Table 2. Since these are less frequent words, initializing label embeddings as word embeddings yields a worst result than learning label embeddings from scratch.

1.3.3. Comparing to a SOTA Single-Label Denoise Method

While there is no regularization-based unsupervised method developed for noisy single-label learning that can be directly applied to MLC, we modify *MD-DYR-SH* (Arazo et al., 2019) so that we can see how a SOTA single-label denoise method performs compared to CbMLC. MD-DYR-SH is based on *mixup* (Hongyi Zhang, 2018), which assumes inputs to be real numbers. For this reason, MD-DYR-SH can only be applied on *Pascal VOC*.

MD-DYR-SH leverages an observation that a deep model fits clean labels before noisy labels: they fit training loss

[1] Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

Dataset	Input Type	#Train	#Val	#Test	Labels (L)	Features	Mean Label /Sample	Median Label /Sample	Max Label /Sample	Mean Sample /Label	Median Label /Sample	Max Label /Sample
reuters	Sequential	6,993	777	3,019	90	23662	1.23	1	15	106.50	18	2,877
rcv1	Sequential	703,135	78,126	23,149	103	368,998	3.21	3	17	24,362	7,250	363,991
bibtex	Binary Vector	4,377	487	2515	159	1836	2.38	2	28	72.79	54	689
delicious	Binary Vector	11,579	1,289	3,185	983	500	19.06	20	25	250.15	85	5,189
Pascal VOC	Image	5011	N/A	4952	20	448 × 448	1.56	1	7	395.65	268	2,095

Table 1. **Dataset Statistics**. We select five datasets that vary in a number of aspects including the input type, number of samples, number of labels, and how imbalanced the label classes are.

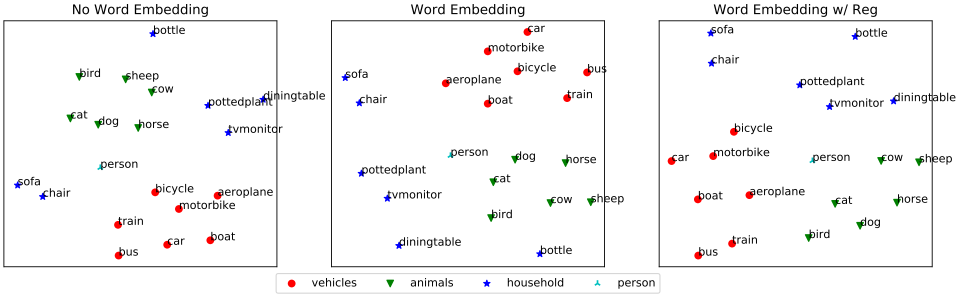

Figure 1. t-SNE visualization of various label embeddings on Pascal VOC. **Left:** Randomly initialized label embeddings learned over the training process. **Middle:** Fixed word embeddings as label embeddings. **Right:** Label embeddings initialized as word embeddings and learned with context-based regularization during training. *While the first setting already produces label embeddings with good qualities, the last setting learns "perfect" label clusters.*

	No WordEmb	WordEmb 0.01	WordEmb 0.10	WordEmb 1.00
miF1	***0.4119***	0.4117	0.4106	0.4104
maF1	***0.2538***	0.238	0.2375	0.2354

Table 2. Performance comparison of using randomly initialized label embeddings to using label embbedings with context-based regularization on the *ebird* dataset. The numbers next to WordEmb are the regularization coefficients. Highlighted numbers are the best of a row.

F1-score	**MD-DYR-SH**	**ML-GCN**	**CbMLC**
miF1	0.7290	0.7138	***0.7621***
maF1	0.6396	0.6281	***0.7281***

Table 3. Comparison of CbMLC to ML-GCN, a SOTA multi-label image recognition method, and MD-DYR-SH, a SOTA single-label denoise method. Highlighted numbers are the best of a row.

curve with a beta mixture models (BMM) to identify noisy labels. To adapt it for MLC, instead of only penalizing false negatives, we also add loss for false positives. We test MD-DYR-SH with the combined noise. We fine tune MD-DYR-SH to achieve its best possible performance and show the results in Table 3. We can see that although, modified MD-DYR-SH slightly improves over ML-GCN, a SOTA multi-label image recognition method, the gap between MD-DYR-SH and CbMLC is still large. Our intepretation is that single-label learning only optimizes on the one true positive label per instance, but the MLC loss simultaneously penalizes all misclassified labels for an instance; therefore, we need to develop regularization techniques that is suitable for the MLC loss rather than simply adapting approaches from single-label learning.

1.3.4. Complete Results for Type 1 and Type 2 Noise

We add the performance comparison of CbMLC to the best SOTA model on *delicious* and *rcv1*. We present the results for type 1 noise in Table 4 and the results for type 2 noise in Table 5. For type 1 noise, the best SOTA model outperforms CbMLC at the beginning on *delicious*, and the performance gap shrinks as the noise becomes more severe. For *rcv1*, our method generally improves over the best SOTA model. However, due to the large numbers of labels in these two datasets, the noise we injected is relatively small, and therefore CbMLC provides a limited improvement.

For type 2 noise on *delicious*, better maF1 scores from CbMLC indicate that we are better with handling imbalanced classes, and better miF1 scores from the best SOTA model suggests that it performs better for classes with frequent labels under this noisy setting. For type 2 noise on *rcv1*, again, our improvement is larger for classes with less frequent labels. Further, we can also improve on miF1 scores when the noise becomes more severe.

Noise Level		**0.01**		**0.02**		**0.03**		**0.04**		**0.05**	
Dataset	**Method**	**miF1**	**maF1**	**miF1**	**maF1**	**miF1**	**maF1**	**miF1**	**maF1**	**miF1**	**maF1**
delicious	**SOTA**	***0.3794***	***0.1734***	***0.3732***	***0.1605***	***0.3677***	***0.1532***	***0.3612***	***0.1476***	***0.3452***	0.1352
	CbMLC	0.3668	0.1409	0.3544	0.1409	0.3543	0.1409	0.3482	0.1409	0.3306	***0.1409***
rcv1	**SOTA**	0.8720	0.7265	***0.8857***	0.7232	***0.8704***	0.7123	0.8704	0.7021	***0.8721***	0.7158
	CbMLC	***0.8725***	***0.7372***	0.8726	***0.7249***	0.8688	***0.7242***	***0.8706***	***0.7212***	0.8689	***0.7236***

Table 4. Performance comparison of CbMLC to the best SOTA model on the additional two datasets with type 1 noise. Higher scores are better. Highlighted numbers are the best of in each row.

Noise Level		**0.1**		**0.2**		**0.3**		**0.4**		**0.5**	
Dataset	**Method**	**miF1**	**maF1**	**miF1**	**maF1**	**miF1**	**maF1**	**miF1**	**maF1**	**miF1**	**maF1**
delicious	**SOTA**	***0.3689***	0.1871	***0.3646***	0.1867	***0.3644***	0.1784	***0.3634***	0.1755	***0.3620***	0.1652
	CbMLC	0.3668	***0.1933***	0.3544	***0.1911***	0.3543	***0.1839***	0.3482	***0.1763***	0.3306	***0.1708***
rcv1	**SOTA**	0.8711	***0.7374***	0.8707	***0.7357***	0.8648	0.7147	0.8630	0.7163	0.8647	0.7166
	CbMLC	***0.8720***	0.7430	***0.8710***	0.7350	***0.8712***	***0.7355***	***0.8674***	***0.7393***	***0.8658***	***0.7290***

Table 5. Performance comparison of CbMLC to the best SOTA model on the additional two datasets with type 2 noise. Higher scores are better. Highlighted numbers are the best of in each row.

References

Arazo, E., Ortego, D., Albert, P., O'Connor, N., and McGuinness, K. Unsupervised label noise modeling and loss correction. In *International Conference on Machine Learning*, pp. 312–321. PMLR, 2019.

Hongyi Zhang, Moustapha Cisse, Y. N. D. D. L.-P. mixup: Beyond empirical risk minimization. *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=r1Ddp1-Rb`.

Sullivan, B. L., Wood, C. L., Iliff, M. J., Bonney, R. E., Fink, D., and Kelling, S. ebird: A citizen-based bird observation network in the biological sciences. *Biological conservation*, 142(10):2282–2292, 2009.

Van der Maaten, L. and Hinton, G. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.