Exploiting Conjugate Label Information for Multi-Instance Partial-Label Learning (Appendix)

Wei Tang[1,2], **Weijia Zhang**[3], **Min-Ling Zhang**[1,2*]
[1]School of Computer Science and Engineering, Southeast University, Nanjing 210096, China
[2]Key Lab. of Computer Network and Information Integration (Southeast University), MoE, China
[3]School of Information and Physical Sciences, The University of Newcastle, NSW 2308, Australia
tangw@seu.edu.cn, weijia.zhang@newcastle.edu.au, zhangml@seu.edu.cn

1 Additional Experiment Results

1.1 Results of PLL algorithms with MLP

We compared with four PLL algorithms, i.e., PRODEN [Lv *et al.*, 2020], RC [Feng *et al.*, 2020], LWS [Wen *et al.*, 2021], and PL-AGGD [Wang *et al.*, 2022]. Among these, the first three algorithms can be used with either linear classifiers or multi-layer perceptrons (MLP). Due to space limitations, we presented only the results obtained using the linear classifiers in the main body of the paper. Table A1 and A2 display the results of ELIMIPL and the comparative PLL algorithms with MLP on the benchmark and CRC-MIPL datasets, respectively.

Table A1 clearly illustrates that ELIMIPL consistently outperforms the classification accuracies of the comparative PLL algorithms with MLP. ELIMIPL consistently outperforms the comparative PLL algorithms in all cases. However, when employing MLP, the comparative PLL algorithms did not consistently achieve superior outcomes when compared to themselves using linear classifiers. This is particularly evident when dealing with datasets containing relatively simple features, where MLP results in lower performance than linear classifiers. This phenomenon suggests that for the comparative PLL algorithms, linear classifiers possess sufficient capacity to handle relatively simple features, while MLP might lead to overfitting on the benchmark datasets.

Table A2 reveals that ELIMIPL significantly outperforms the comparative PLL algorithms in 20 out of 24 cases while showing inferior performance in 3 cases out of 24. Notably, PLL algorithms utilizing MLP consistently outperform those using linear classifiers across almost all cases. When replacing linear classifiers with MLP, results obtained from the KMeansSeg image bag generator exhibit a substantial improvement compared to those generated by simpler image bag generators (i.e., Row and SBN), while the improvements are less pronounced with the SIFT image bag generator. In conclusion, although the PLL algorithms can attain satisfactory results using MLP in certain scenarios, the development of dedicated MIPL algorithms is essential.

1.2 Win/tie/loss counts of Experimental Results

To ensure the reliability of the results, we perform the pairwise t-test at a significance level of 0.05. We present the win/tie/loss counts between ELIMIPL and the comparative algorithms on the benchmark datasets for varying numbers of false positive labels ($r \in \{1, 2, 3\}$), as well as the CRC-MIPL dataset, in Table A3. Several key observations emerge: (a) ELIMIPL demonstrates statistical superiority over MIPL and PLL algorithms in 67.7% and 96.9% of cases, respectively. (b) Across the benchmark datasets, ELIMIPL exhibits statistical superiority over comparative algorithms in 95.3% of cases. (c) Specifically, for the CRC-MIPL dataset, ELIMIPL

Algorithm	r	MNIST	FMNIST	Birdsong	SIVAL
ELIMIPL	1	**.992±.007**	**.903±.018**	**.771±.018**	**.675±.022**
	2	**.987±.010**	**.845±.026**	**.745±.015**	**.616±.025**
	3	**.748±.144**	**.702±.055**	**.717±.017**	**.600±.029**
Mean					
PRODEN	1	.555±.033	.652±.033	.303±.016	.303±.020
	2	.372±.038	.463±.067	.287±.017	.274±.022
	3	.285±.032	.288±.039	.278±.006	.242±.009
RC	1	.660±.031	.697±.166	.329±.014	.344±.014
	2	.577±.039	.684±.029	.301±.014	.299±.015
	3	.362±.029	.414±.050	.288±.019	.256±.013
LWS	1	.605±.030	.702±.033	.344±.018	.346±.014
	2	.431±.024	.547±.040	.310±.014	.312±.015
	3	.335±.029	.411±.033	.289±.021	.286±.018
MaxMin					
PRODEN	1	.465±.023	.358±.019	.339±.010	.322±.018
	2	.338±.031	.315±.023	.329±.016	.295±.021
	3	.260±.037	.265±.031	.305±.015	.244±.018
RC	1	.518±.022	.421±.016	.379±.014	.304±.015
	2	.462±.028	.363±.018	.359±.015	.268±.023
	3	.366±.039	.294±.053	.332±.024	.244±.014
LWS	1	.457±.028	.346±.033	.349±.013	.345±.013
	2	.351±.043	.323±.031	.336±.013	.314±.019
	3	.274±.037	.267±.034	.307±.016	.268±.019

Table A1: The classification accuracies (mean±std) of ELIMIPL and comparative PLL algorithms on the benchmark datasets with varying numbers of false positive candidate labels ($r \in \{1, 2, 3\}$).

Algorithm	Row	SBN	KMeans	SIFT
ELIMIPL	.433±.008	.509±.007	.546±.012	**.540±.010**
Mean				
PRODEN	.405±.012	.515±.010	.512±.014	.352±.015
RC	.290±.010	.394±.010	.304±.017	.248±.008
LWS	.360±.008	.440±.009	.422±.035	.338±.009
MaxMin				
PRODEN	**.453±.009**	**.529±.010**	**.563±.011**	.294±.008
RC	.347±.013	.432±.008	.366±.010	.204±.008
LWS	.381±.011	.442±.009	.335±.049	.287±.009

Table A2: The classification accuracies (mean±std) of ELIMIPL and comparative PLL algorithms on the real-world datasets.

*Corresponding author

	ELIMIPL against						In total
	DEMIPL	MIPLGP	PRODEN	RC	LWS	PL-AGGD	
$r = 1$	2/2/0	3/1/0	16/0/0	16/0/0	16/0/0	8/0/0	61/3/0
$r = 2$	3/1/0	3/1/0	16/0/0	16/0/0	16/0/0	8/0/0	62/2/0
$r = 3$	2/2/0	2/2/0	16/0/0	16/0/0	16/0/0	8/0/0	60/4/0
CRC-MIPL	4/0/0	2/1/0	11/2/3	16/0/0	16/0/0	6/0/2	55/3/5
In total	11/5/0	10/5/0	59/2/3	64/0/0	64/0/0	30/0/2	**238/12/5**

Table A3: Win/tie/loss counts on the classification performance of ELIMIPL against the comparing algorithms.

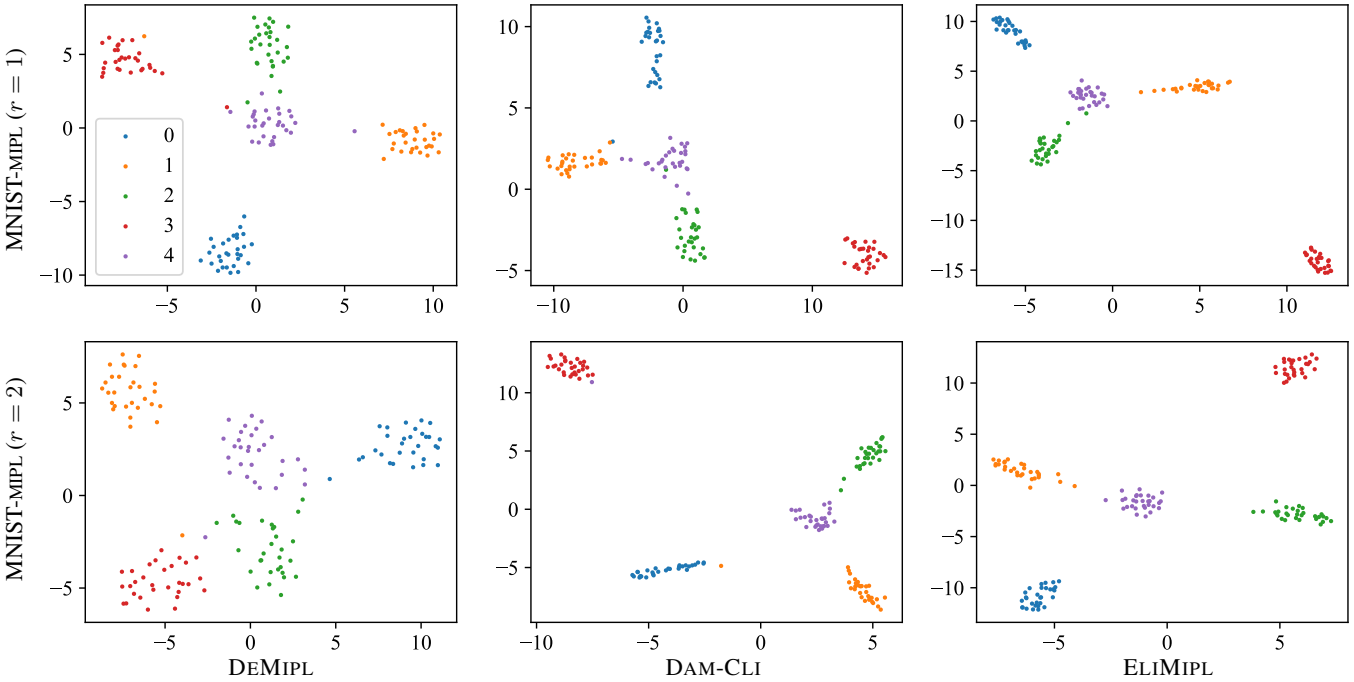


Figure A1: t-SNE visualization of aggregated bag-level feature representations produced by the attention mechanisms in DEMIPL, DAM-CLI, and ELIMIPL on the test set of the MNIST-MIPL dataset ($r \in \{1,\ 2\}$).

Algorithm	Row	SBN	KMeans	SIFT
ELIMIPL	.433±.008	.509±.007	.546±.012	.540±.010
DAM-CLI	.424±.007	.501±.008	.534±.012	.531±.010

Table A4: The classification accuracies of ELIMIPL and DAM-CLI.

shows statistical superiority over the comparative algorithms in 87.3% of cases. In summary, ELIMIPL achieves either superior or competitive performance comparative to the MIPL and PLL algorithms.

1.3 Effectiveness of the Attention Mechanism

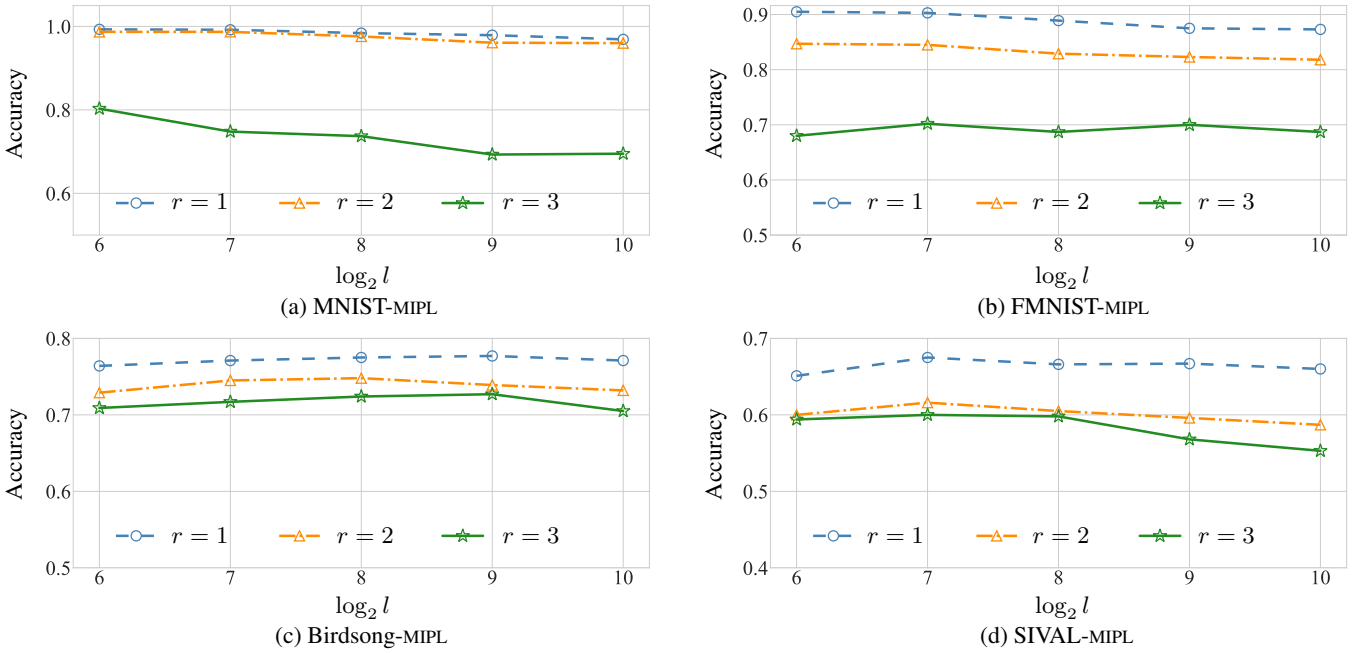
We now show the results to demonstrate the effectiveness of our scaled additive attention mechanism by contrasting it with the disambiguation attention mechanism proposed previously [Tang *et al.*, 2023]. DAM-CLI is derived by substituting the scaled additive attention mechanism in ELIMIPL with the disambiguation attention mechanism from DEMIPL. Consequently, the sole distinction between ELIMIPL and DAM-CLI lies in the utilization of different attention mechanisms.

Table A4 illustrates that ELIMIPL consistently attains higher average accuracies compared to DAM-CLI, indicating the effectiveness of the scaled additive attention mechanism.

1.4 Visualization of the Feature Representations

To delve deeper into the scaled additive attention mechanism, we employ t-SNE [Van der Maaten and Hinton, 2008] to visualize the aggregated bag-level feature representations, i.e., $\boldsymbol{z}_i$ in Eq. (4), on the test set of the MNIST-MIPL dataset when $r \in \{1, 2\}$. The t-SNE algorithm is implemented by the `sklearn.manifold` package with default parameters.

Figure A1 illustrates the feature representations generated by the attention mechanisms in DEMIPL, DAM-CLI, and ELIMIPL on the test set of the MNIST-MIPL dataset when $r \in \{1, 2\}$. Here, DAM-CLI signifies the use of the disambiguation attention mechanism in DEMIPL to replace the scaled additive attention mechanism in ELIMIPL. In Figure A1, the feature representations produced by the disambiguation attention mechanism in DEMIPL exhibit more intersections between different categories, suggesting the reduced discriminations of the representations. In contrast, the feature representations generated by the disambiguation attention mechanisms in DAM-CLI form more compact clusters than those produced by the disambiguation attention mechanisms in DEMIPL. Additionally, the feature representations generated by the scaled additive attention mechanisms in ELIMIPL exhibit increased accuracy and separability compared to those produced by the disambiguation attention mechanisms



(a) MNIST-MIPL (b) FMNIST-MIPL (c) Birdsong-MIPL (d) SIVAL-MIPL

Figure A2: The classification accuracies of ELIMIPL with varying l on the benchmark datasets.

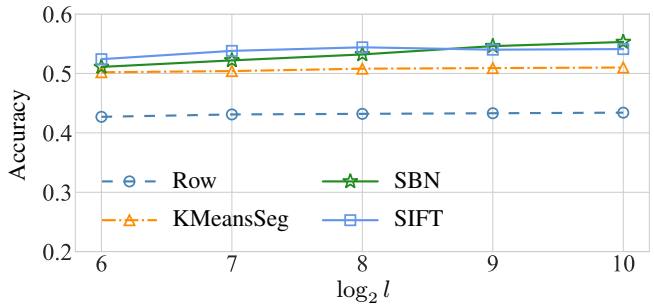


Figure A3: The classification accuracies of ELIMIPL with varying l on the CRC-MIPL dataset.

in DEMIPL and DAM-CLI.

Consequently, the proposed scaled additive attention mechanism is more effective for aggregating bag-level feature representations than the disambiguation attention mechanism proposed in DEMIPL [Tang *et al.*, 2023]. Moreover, the CLI enhances the attention mechanism's ability to aggregate more compact bag-level feature representations.

1.5 Robustness to the Embedded Space Dimension

In the instance-level feature extractor, the feature transformation network $\psi_2(\cdot)$ maps instance-level features to an embedded space of dimension l. Furthermore, the scaling factor in the scaled additive attention is $1/\sqrt{l}$. To examine the impact of parameter l on disambiguation outcomes, we vary l within the set $\{64, 128, 256, 512, 1024\}$. Especially, for each dataset, all experiments maintain consistent data partitioning and other parameters except for varying l.

Figures A2 and A3 depict the classification accuracies of ELIMIPL on the benchmark and CRC-MIPL datasets with varying l, respectively. On the benchmark dataset, ELIMIPL's performance demonstrates insensitivity to dimension l for $r = 1$ or 2. However, when $r = 3$, some variations in classification accuracy emerge for the MNIST-MIPL and SIVAL-MIPL datasets. Specifically, an increase in l correlates with a decrease in classification accuracy. This phenomenon can be attributed to the relatively straightforward features of the benchmark datasets, resulting in an undue emphasis on their feature representation when projected into a higher-dimensional embedding space. This effect becomes particularly pronounced under challenging disambiguation conditions, i.e., $r = 3$. Overall, ELIMIPL achieves improved classification accuracies on the benchmark datasets when employing smaller values of dimension l. On the CRC-MIPL dataset, ELIMIPL's classification accuracies remain stable across variations in l. Notably, when utilizing the Row and KMeansSeg image bag generators, ELIMIPL demonstrates strong robustness to the varying dimension l.

Based on the insights derived from the above analysis, we opt for uniform settings of $l = 128$ for the benchmark datasets and $l = 512$ for the CRC-MIPL dataset, corresponding to $\log_2 l = 7$ and $\log_2 l = 9$, respectively. From Figures A2 and A3, it is evident that such parameter configurations of dimension l can yield commendable results.

1.6 Effectiveness of the Scaling Factor

The scaling factor prevents the softmax function from entering regions with small gradients, thereby mitigating the issue of gradient vanishing. To assess the impact of the scaling factor, we introduce a variant called ELIMIPL *wo* $\frac{1}{\sqrt{l}}$, differing from ELIMIPL only by excluding the scaling factor.

Table A5 displays the classification accuracies of ELIMIPL *wo* $\frac{1}{\sqrt{l}}$ and ELIMIPL on the benchmark datasets. The results

Algorithm	r	MNIST	FMNIST	Birdsong	SIVAL
ELIMIPL	1	.992±.007	.903±.018	.771±.018	.675±.022
	2	.987±.010	.845±.026	.745±.015	.616±.025
	3	.748±.144	.702±.055	.717±.017	.600±.029
ELIMIPL *wo* $\frac{1}{\sqrt{l}}$	1	.200±.000	.275±.080	.138±.019	.143±.030
	2	.211±.032	.229±.049	.138±.020	.139±.029
	3	–	.200±.000	.129±.024	.131±.022

Table A5: The classification accuracies (mean±std) of ELIMIPL and comparative algorithms on the benchmark datasets with varying numbers of false positive candidate labels ($r \in \{1, 2, 3\}$). The symbol "–" indicates that ELIMIPL *wo* $\frac{1}{\sqrt{l}}$ fails to achieve accuracy due to gradient vanishing.

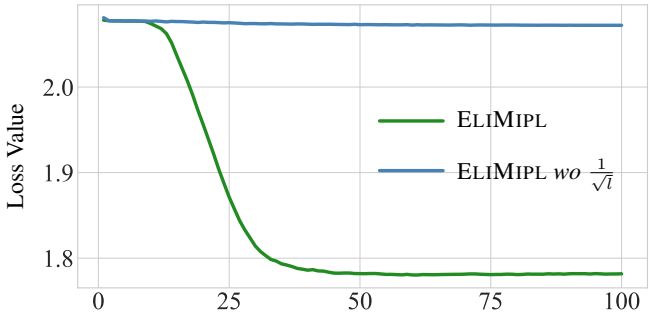


Figure A4: The loss values of ELIMIPL and ELIMIPL *wo* $\frac{1}{\sqrt{l}}$ on the MNIST-MIPL dataset ($r = 1$).

highlight that without the scaling factor, the model's performance resembles random classification, demonstrating ineffective learning and susceptibility to gradient vanishing. Furthermore, Figure A4 illustrates that without the scaling factor, the loss value fails to converge. Therefore, in ELIMIPL, the scaling factor plays a pivotal role in achieving convergence.

1.7 Parameter Sensitivity

Figure A5 illustrates the classification accuracies of ELIMIPL on the FMNIST-MIPL dataset across varying parameters μ and γ. Specifically, μ and γ are chosen from the sets $\{0.7, 0.8, 0.9, 1.0, 1.1\}$ and $\{0.4, 0.5, 0.6, 0.7, 0.8\}$, respectively. ELIMIPL exhibits robustness to various combinations of parameters μ and γ. The accuracies remain stable in most cases, even when $r = 3$. In our experiments, we set $\mu = 1$ and $\gamma = 0.5$ for ELIMIPL on the FMNIST-MIPL dataset. The results of these experiments validate the efficacy of such parameter configurations.

1.8 Computational Complexity

Table A6 presents the floating-point operations (FLOPs), number of parameters (Params), peak GPU memory usage (PM), average time per test multi-instance bag (Times), and average accuracy (Acc) over 10 trials, providing comprehensive metrics to assess model complexity. The complexity of MIPLGP is denoted as $\mathcal{O}((k+1)n^2)$, where k and n represent the number of classes and instances, respectively.

As observed in Table A6, the computational complexities of ELIMIPL and DEMIPL are comparable, yet ELIMIPL exhibits higher accuracy than DEMIPL. Moreover, ELIMIPL achieves this superior accuracy with a lighter computational burden compared to MIPLGP. This suggests that, while keeping the computational cost comparable to that of DEMIPL, ELIMIPL attains superior accuracy, outperforming MIPLGP in terms of both accuracy and computational complexity.

Algorithm	FLOPs (M)	Params (M)	PM (MiB)	Times (s)	Acc
ELIMIPL	109.86	0.43	1824	1.554	.992
DEMIPL	109.86	0.43	1822	1.426	.976
MIPLGP	–	–	12938	1.187	.949

Table A6: The outcomes on the MNIST-MIPL dataset ($r = 1$).

2 Why ELIMIPL Works?

The experimental results presented in the main body of the paper and the supplementary material demonstrate that ELIMIPL outperforms the comparative MIPL and PLL algorithms across the majority of scenarios. Furthermore, we conduct a thorough validation of the efficacy of each component within ELIMIPL. In this section, we provide insights into the key factors contributing to the success of ELIMIPL.

Scaled Additive Attention Mechanism The t-SNE visualization in Figure A1 reveals that our proposed scaled additive attention mechanism generates feature representations that are not only more compact but also more accurate compared to the disambiguation attention mechanism in DEMIPL [Tang *et al.*, 2023]. The results presented in Table A4 further affirm that the utilization of the scaled additive attention mechanism leads to higher classification accuracy than employing the disambiguation attention mechanism. Additionally, as illustrated in Table A5 and Figure A4, the inclusion of the scaling factor in the scaled additive attention mechanism ensures the model's convergence with satisfactory accuracy. Without the scaling factor, the model fails to converge, resulting in classification outputs resembling random guesses.

CLI loss During training, ELIMIPL learns conjugate label information by minimizing the CLI loss, comprising mapping loss, sparse loss, and inhibition loss. Figure 2 in the main body of the paper illustrates that using CLI loss enhances the predicted probabilities of the classifier on true labels while suppressing probabilities on non-candidate labels. Tables 4 and 5 in the main body of the paper demonstrate that the CLI loss significantly improves the model's performance compared to using mapping loss, mapping loss with sparse loss, mapping loss with inhibition loss, and cross-entropy loss. Thus, CLI loss effectively exploits the information from the label sets and the intrinsic properties of the label space, enhancing the model's disambiguation performance.

In summary, the effectiveness of ELIMIPL can be attributed to two pivotal components: (a) the scaled additive attention mechanism, which is responsible for generating discriminative bag-level feature representations, and (b) CLI loss, which is proficient in exploiting the information from the label sets and the intrinsic properties of the label space.

3 MIPL Datasets

We employ four benchmark datasets and one real-world dataset, all of which are publicly accessible [Tang *et al.*, 2024, 2023]. Next, we will provide detailed descriptions of the data preparation procedures for each of these datasets.

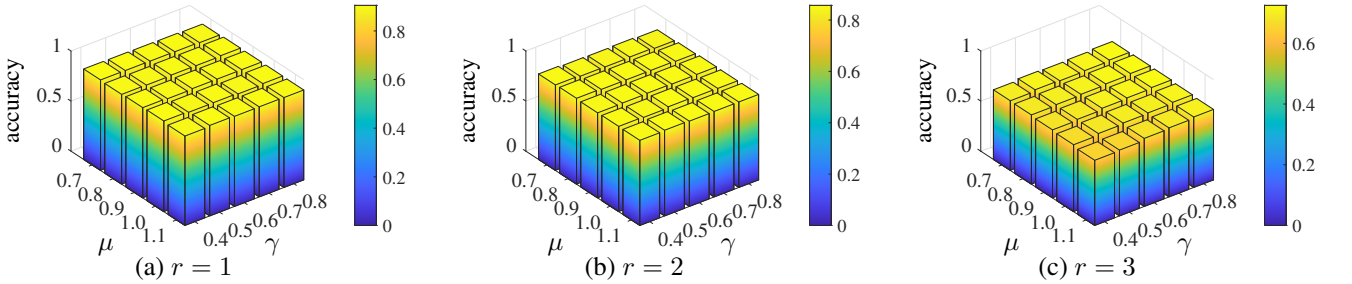


Figure A5: The performance of ELIMIPL with varying μ and γ on the FMNIST-MIPL dataset ($r \in \{1, 2, 3\}$).

3.1 Benchmark MIPL Datasets

The MNIST-MIPL and FMNIST-MIPL datasets are adaptations of the original MNIST and Fashion-MNIST datasets [LeCun *et al.*, 1998; Xiao *et al.*, 2017], respectively. To construct the MNIST-MIPL and FMNIST-MIPL datasets, positive and negative instances within each multi-instance bag are selected from targeted and reserved class labels, respectively. For the MNIST-MIPL dataset, the classes $\{0, 2, 4, 6, 8\}$ are designated as the targeted classes, ensuring the presence of positive instances corresponding to these classes. Conversely, negative instances are drawn randomly from the reserved classes $\{1, 3, 5, 7, 9\}$. Similarly, the FMNIST-MIPL dataset is constructed with the targeted class labels {*T-shirt, Trouser, Coat, Sneaker, Bag*} and the reserved class labels {*Pullover, Dress, Sandal, Shirt, Ankle boot*}.

The Birdsong dataset is widely employed in both multi-instance multi-label learning [Briggs *et al.*, 2012] and PLL [Lv *et al.*, 2020]. This dataset comprises 548 multi-instance bags, collectively containing 10232 instances. Each instance is represented by a 38-dimensional feature vector and corresponds to a single label, which is either one of the 13 specific target classes or a singular negative class. In the Birdsong-MIPL dataset, the 13 targeted classes are utilized to select positive instances, while the negative class serves as the reserved label encompassing the negative instances.

The SIVAL is a MIL benchmark dataset for content-based image retrieval with 1500 images [Settles *et al.*, 2007]. Each image serves as a multi-instance bag containing either 31 or 32 instances, linked to one of 25 distinct class labels. Each instance is characterized by a feature vector in a 30-dimensional space. To create the SIVAL-MIPL dataset from the SIVAL dataset, the arrangement of multi-instance bags remains unchanged. Every candidate label set is generated by retaining the true label and randomly choosing r false positive labels from the remaining 24 classes.

3.2 Real-World MIPL Datasets

The CRC-MIPL dataset consists of 7000 images used for classifying colorectal cancer in the absence of exact labels. These images are uniformly selected from the 7 classes of the NCT-CRC-HE-100K dataset [Kather *et al.*, 2019]. To form a candidate label set for each image, an expert trains three crowdsourced workers before annotation. The final candidate label set is obtained by aggregating the candidate labels from all three workers. The methodology is elaborated as follows: Firstly, workers assign candidate labels with non-zero probabilities, thereby creating a label set per image. Higher probabilities indicate a greater likelihood of being the true label, whereas zero probabilities indicate non-candidate labels. Secondly, the aggregated candidate label set is derived from the three label sets, which includes labels present in two or three sets. If the aggregated set contains only one or no label, the labels with the highest probabilities in each set are selected. In contrast to requiring expert annotation of true labels for each image, this annotation approach effectively reduces the expert workload while achieving satisfactory outcomes.

4 The Image Bag Generators

To learn multi-instance features on the CRC-MIPL dataset, we employ four image bag generators [Wei and Zhou, 2016]:

Row Generator [Maron and Ratan, 1998]: This approach treats each row within the image as an independent instance. For feature extraction, it calculates the average RGB color value of each row and analyzes the color differences with adjacent rows. The resulting instance feature encompasses the RGB values of the current instance, along with the disparities in RGB values between the current instance and the preceding one, as well as the subsequent one. This procedure yields a 9-dimensional feature representation for each instance.

SBN Generator [Maron and Ratan, 1998]: This approach utilizes five 2×2 blobs within the image to generate an instance-level feature. This feature includes RGB color values of the central blob and its four neighboring blobs. Instances are generated by iteratively shifting one pixel at a time, while the SBN generator omits feature information at the image's four corners. This results in a 15-dimensional feature vector for each instance.

KMeansSeg Generator [Zhang *et al.*, 2002]: This generator partitions the image into k segments, producing 6-dimensional features for each segment. The initial three dimensions represent color values within the YCbCr color space, while the subsequent three dimensions derived through wavelet transformation of the luminance Y component.

SIFT Generator [Lowe, 2004]: Using the scale-invariant feature transform (SIFT) algorithm, the SIFT generator divides instances into multiple 4×4 subregions and maps gradients of pixels within these subregions to 8 bins. As a result, SIFT generates a 128-dimensional feature for each instance.

References

Forrest Briggs, Xiaoli Z. Fern, and Raviv Raich. Rank-loss support instance machines for MIML instance annotation.

In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Beijing, China*, pages 534–542, 2012.

Lei Feng, Jiaqi Lv, Bo Han, Miao Xu, Gang Niu, Xin Geng, Bo An, and Masashi Sugiyama. Provably consistent partial-label learning. In *Advances in Neural Information Processing Systems 33, Virtual Event*, pages 10948–10960, 2020.

Jakob Nikolas Kather, Johannes Krisam, Pornpimol Charoentong, Tom Luedde, Esther Herpel, Cleo-Aron Weis, Timo Gaiser, Alexander Marx, Nektarios A Valous, Dyke Ferber, et al. Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. *PLoS Medicine*, 16(1):e1002730:1–22, 2019.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

Jiaqi Lv, Miao Xu, Lei Feng, Gang Niu, Xin Geng, and Masashi Sugiyama. Progressive identification of true labels for partial-label learning. In *Proceedings of the 37th International Conference on Machine Learning, Virtual Event*, pages 6500–6510, 2020.

Oded Maron and Aparna Lakshmi Ratan. Multiple-instance learning for natural scene classification. In *Proceedings of the 15th International Conference on Machine Learning, Madison, Wisconsin, USA*, pages 341–349, 1998.

Burr Settles, Mark Craven, and Soumya Ray. Multiple-instance active learning. In *Advances in Neural Information Processing Systems 20, Vancouver, British Columbia, Canada*, pages 1289–1296, 2007.

Wei Tang, Weijia Zhang, and Min-Ling Zhang. Disambiguated attention embedding for multi-instance partial-label learning. In *Advances in Neural Information Processing Systems 36, New Orleans, LA, USA*, pages 56756–56771, 2023.

Wei Tang, Weijia Zhang, and Min-Ling Zhang. Multi-instance partial-label learning: Towards exploiting dual inexact supervision. *Science China Information Sciences*, 67(3):Article 132103: 1–14, 2024.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(11), 2008.

Deng-Bao Wang, Min-Ling Zhang, and Li Li. Adaptive graph guided disambiguation for partial label learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):8796–8811, 2022.

Xiu-Shen Wei and Zhi-Hua Zhou. An empirical study on image bag generators for multi-instance learning. *Machine Learning*, 105:155–198, 2016.

Hongwei Wen, Jingyi Cui, Hanyuan Hang, Jiabin Liu, Yisen Wang, and Zhouchen Lin. Leveraged weighted loss for partial label learning. In *Proceedings of the 38th International Conference on Machine Learning, Virtual Event*, pages 11091–11100, 2021.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms. *CoRR*, abs/1708.07747, 2017.

Qi Zhang, Sally A. Goldman, Wei Yu, and Jason E. Fritts. Content-based image retrieval using multiple-instance learning. In *Proceedings of the 19th International Conference on Machine Learning, Sydney, Australia*, pages 682–689, 2002.