Multiple Instance Learning with Bag Dissimilarities

Veronika Cheplygina, David M.J. Tax, Marco Loog

Bag similarity network for deep multi-instance learning

Xinggang Wang, Yongluan Yan, Peng Tang, Wenyu Liu, Xiaojie Guo

Presenter: Qingmei Wang

Outline

- Background
- Bag Dissimilarity Representation
- Bag Similarity Network
- Summary

Background

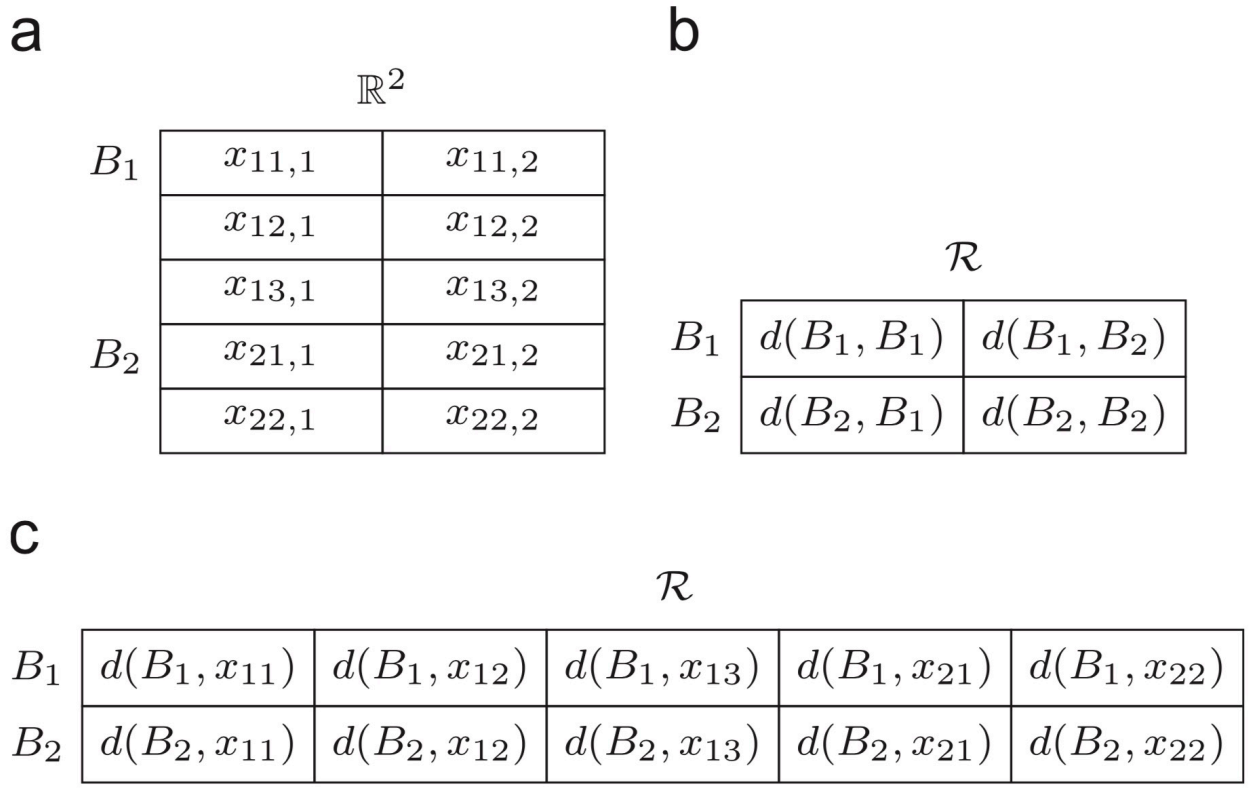

Fig. 1. Representations of a MIL problem with 2 bags and 2 features. B_1 has 3 instances and B_2 has 2 instances. The dimensionality of the original representation depends on the number of features, while in the dissimilarity representation, the dimensionality depends on the number of bags or instances. (a) Original MIL, (b) bag dissimilarity and (c) instance dissimilarity.

In MIL, each sample is in the form of a labeled bag that contains a set of instances associated with input features.

In MIL, the assumption is that a positive bag contains at least one positive instance, whereas a negative bag contains only negative instances.

The goal of MIL in a binary task is to train a classifier so that the labels of testing bags may be predicted.

Background

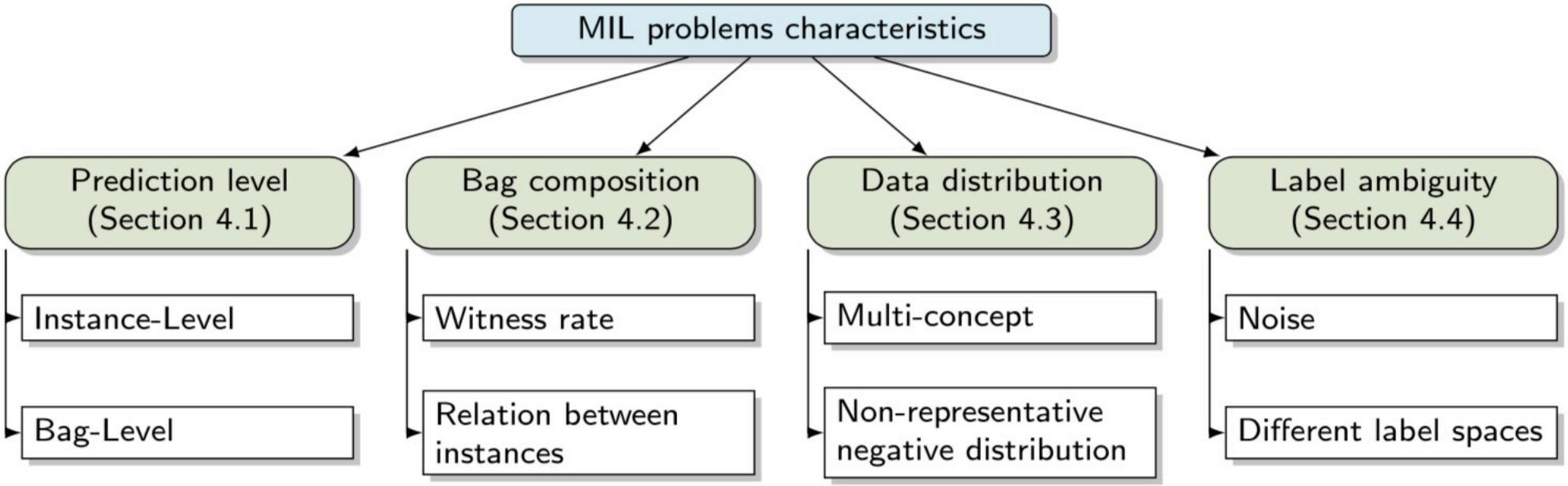

Fig. 1. Characteristics inherent to MIL problems.

Witness rate : the scenario where the number of positive instances in a bag

Bag Dissimilarity Representation

Represent a MIL bag by its dissimilarities to prototype objects in a representation set R .
R is taken to be a subset of size M of the training set of size N (typically M<N) .

$$\mathbf{d}(B_i, \mathcal{T}) = [d(B_i, B_1), \ldots d(B_i, B_M)]$$

Each bag is represented by a single feature vector.
MIL problem can be viewed as a regular supervised learning problem.

Bag Dissimilarity Representation

Point set

Hausdorff distance applied to bags uses the maximum mismatch between the instances of the respective bags

$$d_h(B_i, B_j) = \max_k \min_l d(\mathbf{x}_{ik}, \mathbf{x}_{jl}) \tag{1}$$

Modified versions of the Hausdorff distance

$$d_{minmin}(B_i, B_j) = \min_k \min_l d(\mathbf{x}_{ik}, \mathbf{x}_{jl}) \text{ and} \tag{2}$$

$$d_{meanmin}(B_i, B_j) = \frac{1}{n_i} \sum_{k=1}^{n_i} \min_l d(\mathbf{x}_{ik}, \mathbf{x}_{jl}). \tag{3}$$

$$d_{meanmean}(B_i, B_j) = \frac{1}{n_i n_j} \sum_{k=1}^{n_i} \sum_{l=1}^{n_j} d(\mathbf{x}_{ik}, \mathbf{x}_{jl}). \tag{4}$$

Bag Dissimilarity Representation

Point set

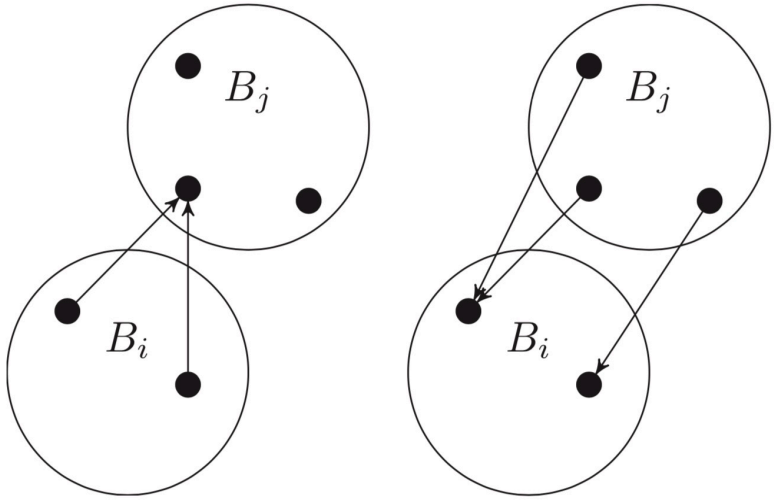

Fig. 2. Minimum instance distances between two bags. The bag dissimilarity is defined as the minimum, maximum, average or other statistic of these distances. The directions of the arrows show that there are two, possibly asymmetric, dissimilarities: that of B_i to B_j, and that of B_j to B_i.

shows the first step in computing dissimilarities between two bags. The arrows in each diagram are the minimum distances between instances of two bags

Bag Dissimilarity Representation

Bags as instance distributions

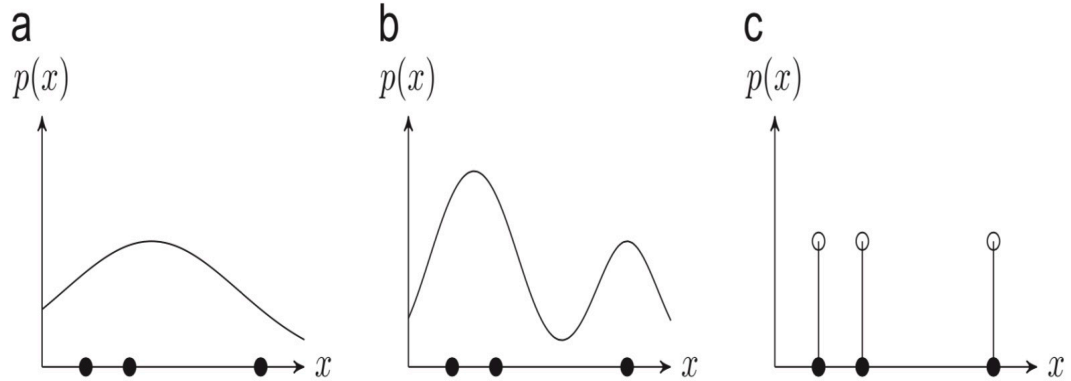

Fig. 3. Different ways to represent a 1-D bag with instances at $x=0.5$, $x=1$ and $x=2.5$ as a distribution. From left to right: normal distribution, Parzen density, δ-peaks. The type of approximation influences the choice of distribution distance that can be applied.

view each bag as a probability distribution in instance space, and define a bag dissimilarity as a distribution distance;
approximate the instance distributions and provide distances between approximated distributions.

a number of approaches to approximate a 1D distribution

Bag Dissimilarity Representation

Bags as instance distributions

Bag dissimilarity measure

$$d_{\text{Maha}}(B_i, B_j) = (\mu_i - \mu_j)^\top \left(\frac{1}{2}\Sigma_i + \frac{1}{2}\Sigma_j\right)^{-1} (\mu_i - \mu_j). \tag{5}$$

$$d_{\text{CS}}(B_i, B_j) = -\log\left(\frac{K_{\sigma_i+\sigma_j}(B_i, B_j)}{(K_{2\sigma_i}(B_i, B_i)K_{2\sigma_j}(B_j, B_j))^{1/2}}\right)$$

where

$$K_\sigma(B_i, B_j) = \sum_{\substack{\mathbf{x}_k \in B_i \\ \mathbf{x}_l \in B_j}} \frac{\exp\left(\frac{1}{-2\sigma^2}(\mathbf{x}_k - \mathbf{x}_l)^\top (\mathbf{x}_k - \mathbf{x}_l)\right)}{(2\pi\sigma^2)^{d/2}}. \tag{6}$$

Bag Dissimilarity Representation

Bags as instance distributions

Bag dissimilarity measure :earth movers distance (EMD)
EMD measures the minimum amount of work to transform one probability into another probability distribution

$$d_{\mathrm{EMD}}(B_i, B_j) = \sum_{\mathbf{x}_k \in B_i, \mathbf{x}_l \in B_j} f(\mathbf{x}_k, \mathbf{x}_l) d(\mathbf{x}_k, \mathbf{x}_l) \tag{7}$$

where $f(\mathbf{x}_k, \mathbf{x}_l)$ is the flow that minimizes the overall distance, and that is subject to constraints that ensure that the only available amounts of earth are transported into available holes, and that all of the earth is indeed transported: $f(\mathbf{x}_k, \mathbf{x}_l) \geq 0$, $\sum_{\mathbf{x}_k \in B_i} f(\mathbf{x}_k, \mathbf{x}_l) \leq 1/n_j$, $\sum_{\mathbf{x}_l \in B_j} f(\mathbf{x}_k, \mathbf{x}_l) \leq 1/n_i$ and $\sum_{\mathbf{x}_k \in B_i,\ \mathbf{x}_l \in B_j} f(\mathbf{x}_k, \mathbf{x}_l) = 1$.

Experiments

Table 2
Point set, symmetrized dissimilarity, logistic and SVM classifiers. AUC and standard error ($\times$ 100), 5 $\times$ 10-fold cross-validation. Bold = best (or not significantly worse) result per dataset.

Classifier	Data	D_{minmin}	$D_{meanmin}$	D_{maxmin}	$D_{meanmean}$
Logistic	C25	**61.4 (2.3)**	54.8 (2.2)	47.5 (2.2)	53.4 (2.1)
	C50	**98.6 (0.4)**	79.6 (1.9)	50.3 (3.3)	65.6 (3.0)
	D25	86.2 (1.8)	**97.8 (0.5)**	96.6 (0.6)	69.4 (2.4)
	D50	91.7 (1.1)	**100.0 (0.0)**	99.6 (0.2)	**100.0 (0.0)**
	M25	54.4 (2.2)	50.8 (1.9)	60.5 (2.3)	**71.4 (1.9)**
	M50	71.5 (2.4)	78.4 (2.1)	**84.3 (1.5)**	69.4 (2.4)
	Musk 1	88.2 (1.8)	**90.2 (1.7)**	**91.8 (1.6)**	84.3 (1.8)
	Musk 2	92.0 (1.2)	**92.6 (1.3)**	**93.3 (1.2)**	82.8 (1.7)
	African	**96.3 (0.4)**	94.4 (0.6)	94.2 (0.6)	91.1 (0.7)
	Ajax	68.4 (1.6)	**98.1 (0.5)**	**97.2 (0.7)**	87.8 (1.1)
	Alt.ath	49.2 (0.8)	**88.5 (1.7)**	83.7 (1.7)	85.2 (1.8)
	BrCr	89.6 (0.6)	**93.6 (0.4)**	91.1 (0.5)	82.3 (0.7)
	Web	69.7 (4.0)	**77.0 (3.2)**	66.8 (3.7)	69.9 (3.3)
LibSVM	C25	**57.7 (2.4)**	52.2 (2.6)	45.9 (2.1)	40.8 (1.5)
	C50	**98.6 (0.4)**	83.9 (2.0)	46.1 (2.4)	66.4 (3.0)
	D25	72.0 (2.4)	**78.9 (2.0)**	**82.7 (1.6)**	68.2 (2.5)
	D50	92.9 (1.1)	**100.0 (0.0)**	99.5 (0.2)	**100.0 (0.0)**
	M25	47.0 (2.2)	50.6 (2.3)	66.2 (2.3)	**71.6 (2.0)**
	M50	72.0 (2.4)	**78.9 (2.0)**	**82.7 (1.6)**	68.2 (2.5)
	Musk 1	92.0 (1.2)	**93.4 (1.2)**	**93.4 (1.3)**	88.2 (1.5)
	Musk 2	94.0 (1.3)	**95.4 (1.4)**	**95.3 (1.2)**	90.3 (1.5)
	African	**96.6 (0.3)**	**96.7 (0.4)**	95.5 (0.5)	90.1 (0.7)
	Ajax	71.1 (1.4)	**98.6 (0.4)**	**97.8 (0.5)**	84.0 (1.2)
	Alt.ath	50.0 (0.0)	**94.9 (1.0)**	91.4 (1.1)	94.2 (1.1)
	BrCr	87.8 (0.6)	**95.5 (0.3)**	92.6 (0.5)	54.4 (2.4)
	Web	53.2 (4.8)	**76.0 (2.7)**	43.3 (3.6)	**77.6 (3.3)**

Experiments

Table 3
Distribution dissimilarities. AUC and standard error (× 100), 5 × 10-fold cross-validation. Bold = best (or not significantly worse) result per dataset.

Classifier	Data	D_{Maha}	D_{EMD}	D_{CS}
Logistic	Musk 1	70.1 (2.3)	**88.7 (1.9)**	84.6 (2.0)
	Musk 2	**91.5 (1.3)**	–	88.8 (1.5)
	African	59.9 (1.5)	**93.3 (0.6)**	85.9 (1.0)
	Ajax	86.9 (1.8)	**97.9 (0.4)**	95.5 (0.7)
	Alt.ath	49.9 (2.5)	**84.0 (1.9)**	59.2 (2.9)
	BrCr	63.3 (1.2)	**94.5 (0.4)**	89.4 (0.8)
	Web	–	69.4 (4.1)	**75.7 (3.4)**
LibSVM	Musk 1	76.5 (2.9)	**89.8 (1.6)**	**88.2 (1.7)**
	Musk 2	**96.0 (0.9)**	–	87.4 (1.8)
	African	64.8 (1.5)	**94.7 (0.4)**	**94.3 (0.5)**
	Ajax	87.3 (1.7)	**98.9 (0.3)**	98.1 (0.3)
	Alt.ath	47.0 (2.3)	**87.4 (1.7)**	41.9 (2.5)
	BrCr	59.7 (1.1)	**95.5 (0.3)**	93.9 (0.4)
	Web	–	**77.7 (2.7)**	69.5 (3.8)

Experiments

Table 4
Properties of dissimilarity matrices: negative eigenfraction (in %), negative eigen-ratio and non-metricity fraction (in %).

Measure	Data	D_{minmin}	$D_{meanmin}$	D_{maxmin}	D_{mean}	D_{Maha}	D_{EMD}	D_{CS}
NEF	Musk 1	31.2	25.5	22.6	21.0	48.9	27.4	27.1
	Musk 2	31.1	25.3	21.7	20.5	28.4	–	96.3
	Afr	47.2	37.8	33.1	29.7	50.0	49.5	35.4
	Ajax	47.9	32.6	41.3	40.0	49.6	29.9	32.5
	Alt	49.5	21.0	22.0	93.1	36.9	37.4	9.2
	BrCr	45.5	30.7	33.6	51.1	49.9	37.3	27.3
	Web	18.6	4.8	3.7	25.9	–	10.3	100.0
NER	Musk 1	0.4	0.2	0.2	0.2	1.0	0.2	0.2
	Musk 2	0.4	0.2	0.2	0.1	0.3	–	28.1
	Afr	0.7	0.4	0.3	0.3	1.0	1.0	0.3
	Ajax	0.5	0.2	0.3	1.1	1.0	0.2	0.3
	Alt	1.0	0.5	0.6	24.1	0.8	0.7	0.8
	BrCr	0.5	0.4	0.3	0.8	1.0	0.6	0.2
	Web	1.0	1.1	0.9	1.1	–	0.5	0.0
NMF	Musk 1	5.5	2.5	1.6	0.7	27.1	3.8	5.5
	Musk 2	7.2	3.2	1.9	0.8	7.9	–	93.3
	Afr	20.1	8.1	8.8	2.3	27.8	11.4	12.4
	Ajax	17.5	0.1	0.5	0.0	18.9	0.3	4.2
	Alt	0.0	0.3	0.9	0.0	1.7	9.7	0.0
	BrCr	7.8	0.8	2.0	4.2	26.7	3.5	13.1
	Web	2.4	0.0	0.0	0.0	–	0.0	97.3

Experiments

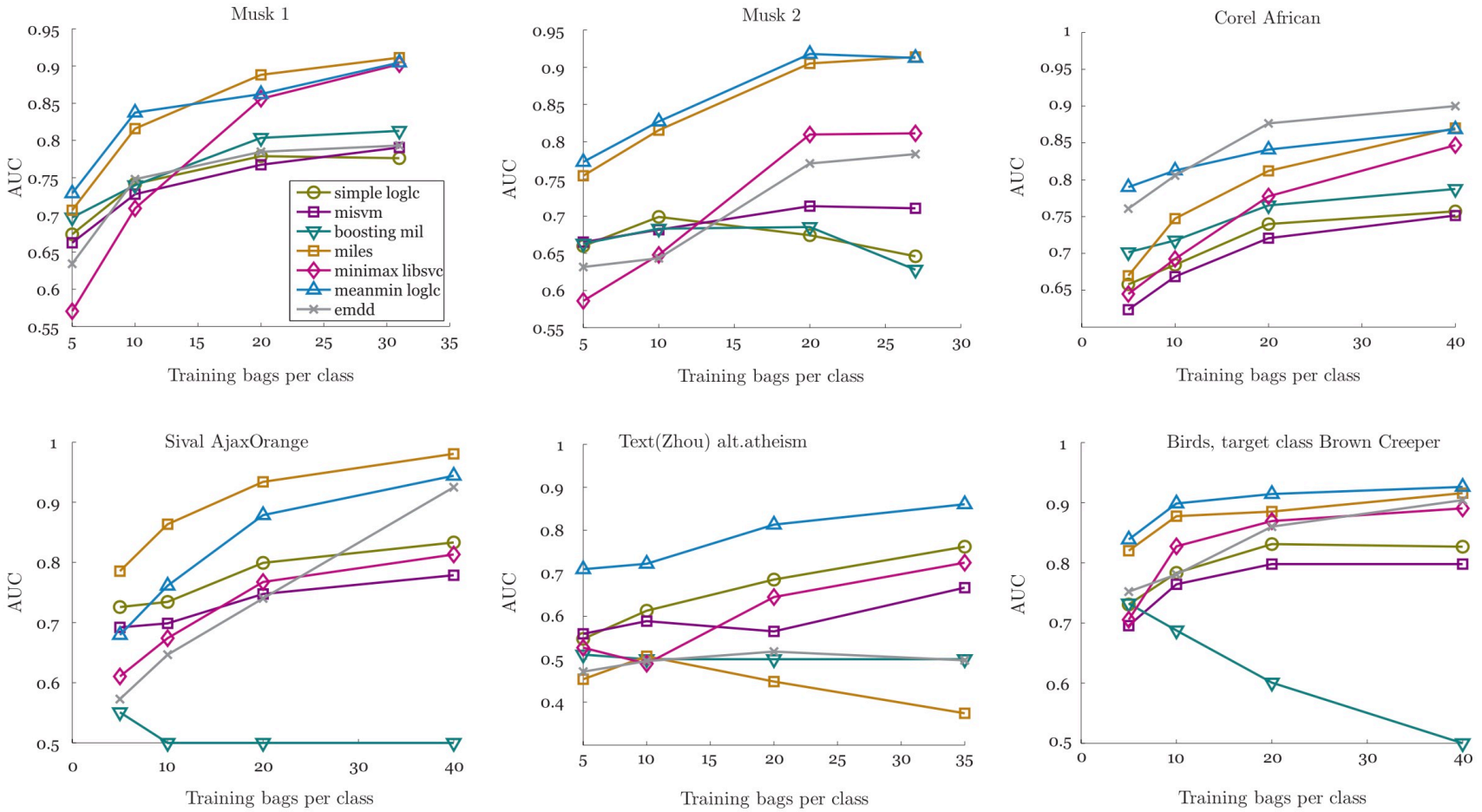

Fig. 7. Learning curves (AUC performance) for Musk 1, Musk 2, African, AjaxOrange, alt.atheism and Brown Creeper datasets. The standard deviations are generally around 0.1, but are lower for MILES, Minimax and MInD at larger training sizes.

Experiments

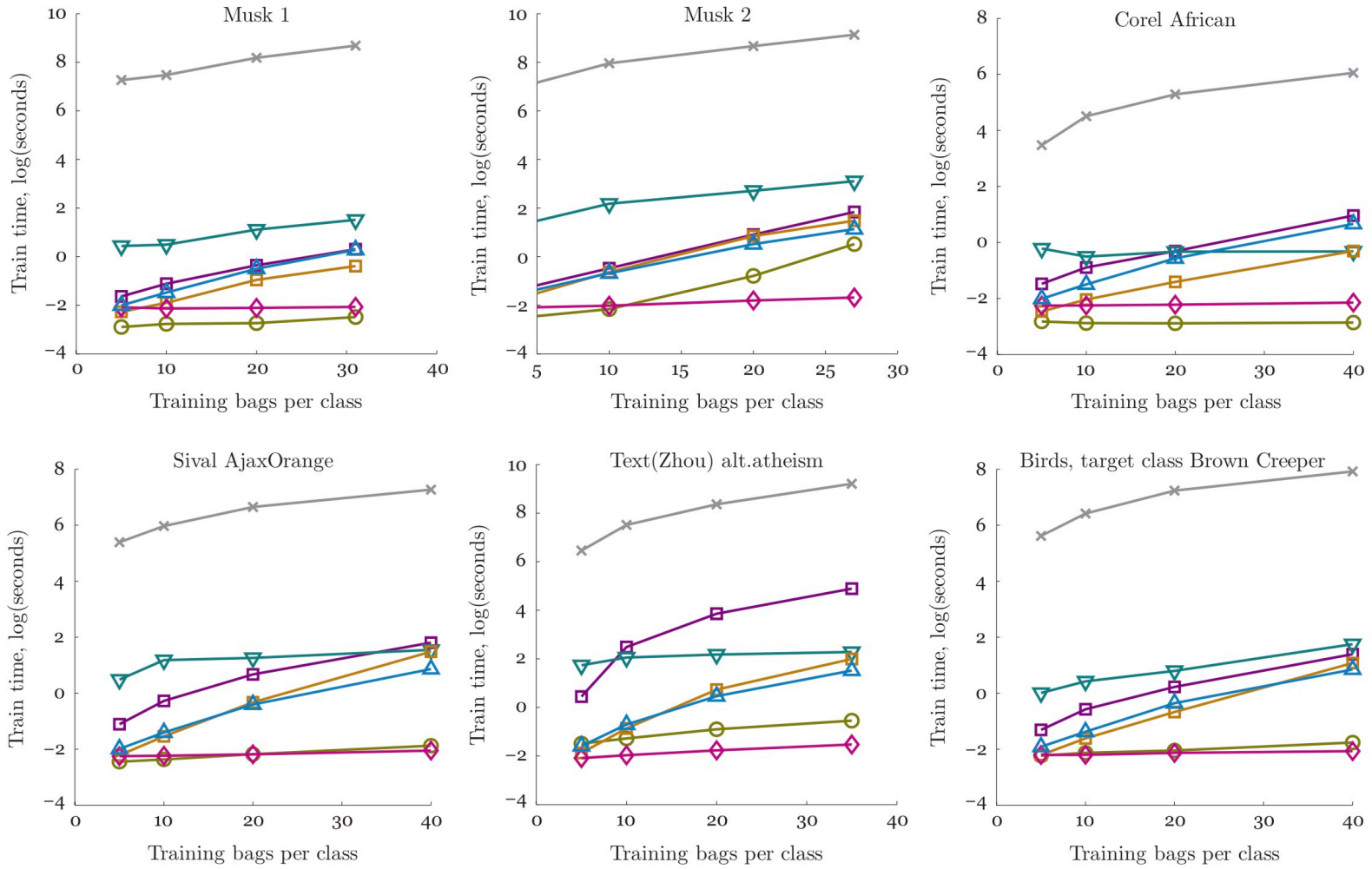

Fig. 8. Training time curves for Musk 1, Musk 2, African, AjaxOrange, alt.atheism and Brown Creeper datasets. The standard deviations in time are all quite low except for EM-DD.

Bag similarity network for MIL

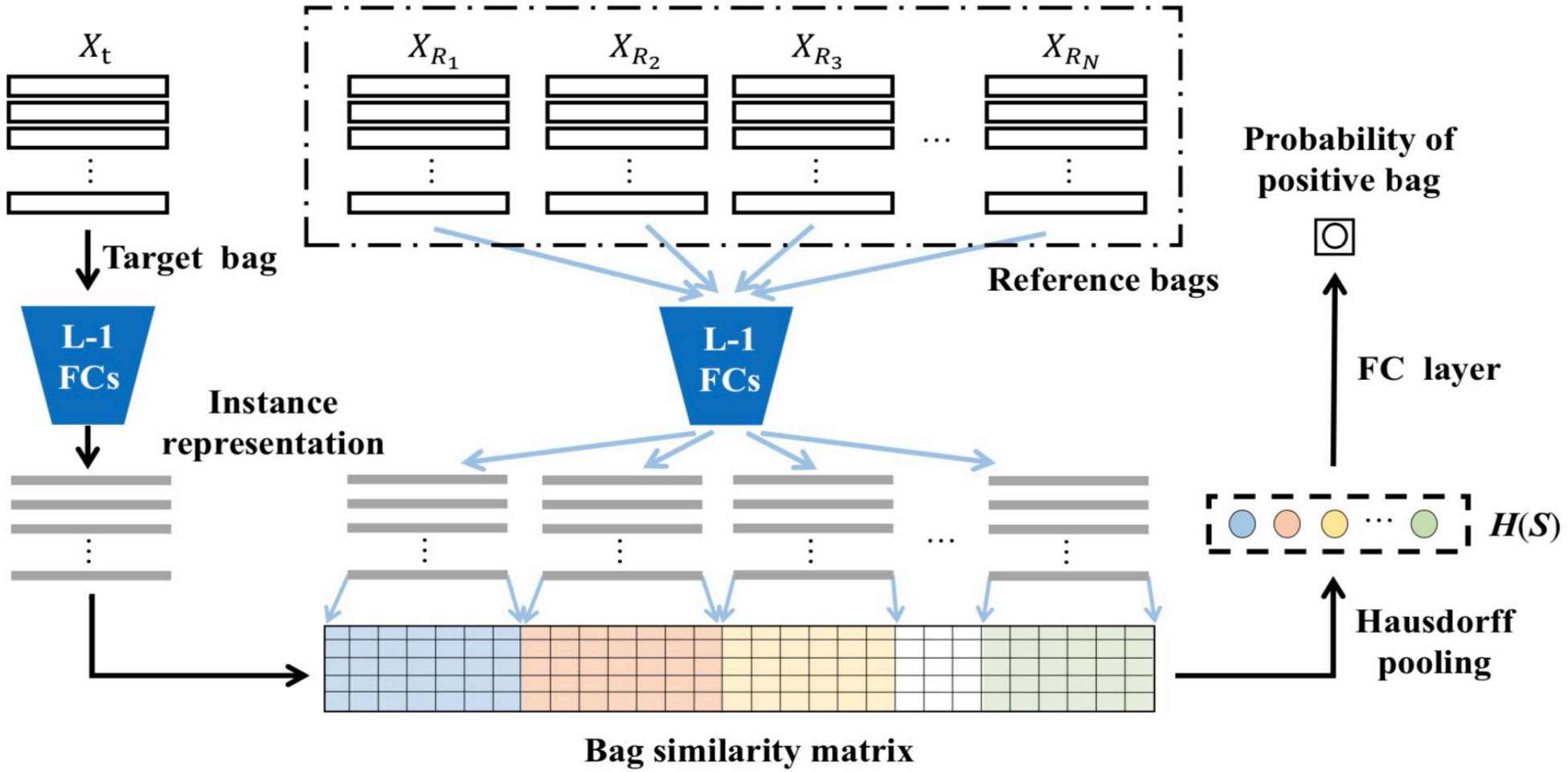

Fig. 2. Architecture of bag similarity network.

Bag similarity network for MIL

Instance-level representations

all training bags $X = \{X_1, X_2, \ldots, X_N\}$ as reference bags

$$X_R^{L-1} = \{X_{R_1}^{L-1}, X_{R_2}^{L-1}, \ldots, X_{R_N}^{L-1}\}, \tag{1}$$

where $X_{R_i}^{L-1} = \{x_{R_i1}^{L-1}, x_{R_i2}^{L-1}, \ldots, x_{R_im_i}^{L-1}\}$ denotes the learned instance-level representations of reference bag X_{R_i}.

Bag similarity network for MIL

Bag similarity matrix

$$S = [f(X_t^{L-1}, X_{R_1}^{L-1}), \ldots, f(X_t^{L-1}, X_{R_N}^{L-1})], \tag{2}$$

where the similarity between X_t and X_{R_i} is represented by $f(X_t^{L-1}, X_{R_i}^{L-1})$, which is defined as

$$\begin{pmatrix} f(x_{t1}^{L-1}, x_{R_i1}^{L-1}) & f(x_{t1}^{L-1}, x_{R_i2}^{L-1}) & \ldots & f(x_{t1}^{L-1}, x_{R_im_i}^{L-1}) \\ f(x_{t2}^{L-1}, x_{R_i1}^{L-1}) & f(x_{t2}^{L-1}, x_{R_i2}^{L-1}) & \ldots & f(x_{t2}^{L-1}, x_{R_im_i}^{L-1}) \\ \ldots & \ldots & \ldots & \ldots \\ f(x_{tm_t}^{L-1}, x_{R_i1}^{L-1}) & f(x_{tm_t}^{L-1}, x_{R_i2}^{L-1}) & \ldots & f(x_{tm_t}^{L-1}, x_{R_im_i}^{L-1}) \end{pmatrix}, \tag{3}$$

where m_t and m_i denote the number of instances in X_t and X_{R_i}, respectively. Therefore, X_t can be represented by a bag similarity matrix, which describes the similarity of X_t to reference bags $X_R = \{X_{R_1}, X_{R_2}, \ldots, X_{R_N}\}$.

Bag similarity network for MIL

Hausdorff pooling

$$H(S) = [h(f(X_t^{L-1}, X_{R_1}^{L-1})), h(f(X_t^{L-1}, X_{R_2}^{L-1})), \ldots, h(f(X_t^{L-1}, X_{R_N}^{L-1}))], \tag{4}$$

where $h(\cdot)$ can be one of the following:

$$\begin{cases} \text{max-max pooling}: & h(A) = \max\limits_i \max\limits_j A_{ij}; \\ \text{mean-max pooling}: & h(A) = \frac{1}{N}\sum_{i=1}^{N} \max\limits_j A_{ij}; \\ \text{min-max pooling}: & h(A) = \min\limits_i \max\limits_j A_{ij}; \\ \text{mean-mean pooling}: & h(A) = \frac{1}{N}\frac{1}{M}\sum_{i=1}^{N}\sum_{j=1}^{M} Aij, \end{cases} \tag{5}$$

Bag similarity network for MIL

Training loss and optimization

To predict bag labels, it is natural to choose the standard cross entropy loss function, which is the same as the loss function in MI-Net

$$L(\hat{Y}_t, Y_t) = -\big((1 - Y_t)\log(1 - \hat{Y}_t) + Y_t \log \hat{Y}_t\big), \tag{6}$$

where $\hat{Y}_t$ is the probability that bag B_t is positive, and Y_t is the label of X_t.

Datasets

Table 1
Parameter details for training BSN with MI-Net, including initial learning rate (LR), weight decay (WD), and the standard deviation of initial weights (W-Std). The parameters for Colon Cancer dataset are the same as in Attention-based MI-Net [13].

Dataset	LR	WD	W-Std
MUSK1	0.001	0.01	0.05
MUSK2	0.001	0.01	0.05
Fox	0.0005	0.01	0.05
Tiger	0.001	0.01	0.05
Elephant	0.0005	0.01	0.05
20 Newsgroups	0.001	0.001	0.1
Messidor	0.0005	0.001	0.05

MUSK1 and **MUSK2** are used to predict the molecular activity of drugs, where bags are molecules and instances are different conformations of these molecules.

Fox, Tiger, and **Elephant** are widely adopted MIL datasets for solving localized content-based image retrieval problems. The bags are images and instances are image segments.

20 Newsgroups contains posts from newsgroups on 20 subjects for text categorization.

Messidor is a public diabetic retinopathy screening dataset that contains 1200 eye fundus images from 654 diabetic and 546 healthy patients.

Colon Cancer contains 100 H&E images generated from normal or malignant tissue appearance. The majority of the nuclei of each cell are marked in each image.

Experiments

Table 2
Bag classification results (*mean* ± *std*) of different methods on MUSK1 and MUSK2 (task: drag activation prediction), as well as Fox, Tiger, and Elephant (task: content-based image retrieval).

Dataset	MUSK1	MUSK2	Fox	Tiger	Elephant
mi-SVM	0.874	0.836	0.582	0.784	0.822
MI-SVM	0.779	0.843	0.578	0.840	0.814
MI-Kernel	0.880	0.893	0.603	0.842	0.843
EM-DD	0.849 ± 0.098	0.869 ± 0.108	0.609 ± 0.101	0.730 ± 0.096	0.771 ± 0.098
mi-Graph	0.889 ± 0.073	*0.903 ± 0.086*	0.616 ± 0.079	*0.860 ± 0.083*	0.869 ± 0.078
miVLAD	0.871 ± 0.097	0.872 ± 0.095	0.620 ± 0.098	0.811 ± 0.087	0.850 ± 0.080
miFV	*0.909 ± 0.089*	0.884 ± 0.094	0.621 ± 0.109	0.813 ± 0.083	0.852 ± 0.081
MInD	0.893 ± 0.019	0.888 ± 0.034	**0.651 ± 0.011**	0.819 ± 0.021	0.857 ± 0.018
MI-Net	0.893 ± 0.099	0.872 ± 0.096	0.627 ± 0.080	0.832 ± 0.087	*0.891 ± 0.074*
Att. Net	0.892 ± 0.040	0.858 ± 0.048	0.615 ± 0.043	0.839 ± 0.022	0.868 ± 0.022
Gated Att. Net	0.900 ± 0.050	0.863 ± 0.042	0.603 ± 0.029	0.845 ± 0.018	0.857 ± 0.027
BSN	**0.931 ± 0.094**	**0.906 ± 0.109**	*0.640 ± 0.111*	**0.878 ± 0.092**	**0.907 ± 0.073**

Table 2 provides the average accuracy and the corresponding standard deviation of methods under comparison.BSN achieves superior performance on MUSK1,MUSK2 ,Tiger and Elephant, and competitive performance on Fox .

Experiments

Table 3
Bag classification results of different methods on 20 Newsgroups (task: text categorization).

Dataset	MI-Kernel	mi-Graph	miFV	MInD	MI-Net	Att. Net	Gated Att. Net	BSN
alt.atheism	0.602 ± 0.039	0.655 ± 0.040	0.848 ± 0.119	*0.861 ± 0.089*	0.846 ± 0.101	0.784 ± 0.084	0.780 ± 0.074	**0.903 ± 0.101**
comp.graphics	0.470 ± 0.033	0.778 ± 0.016	0.594 ± 0.120	0.825 ± 0.118	*0.831 ± 0.123*	0.774 ± 0.081	0.764 ± 0.073	**0.861 ± 0.134**
comp.windows.misc	0.510 ± 0.052	0.631 ± 0.015	0.615 ± 0.172	*0.730 ± 0.094*	*0.730 ± 0.112*	0.686 ± 0.088	0.700 ± 0.080	**0.769 ± 0.101**
comp.ibm.pc.hardware	0.469 ± 0.036	0.595 ± 0.027	0.665 ± 0.147	0.780 ± 0.127	*0.803 ± 0.155*	0.632 ± 0.087	0.640 ± 0.080	**0.813 ± 0.29**
comp.sys.mac.hardware	0.445 ± 0.032	0.617 ± 0.048	0.660 ± 0.157	*0.835 ± 0.098*	0.811 ± 0.138	0.744 ± 0.084	0.754 ± 0.082	**0.865 ± 0.113**
comp.window.x	0.508 ± 0.043	0.698 ± 0.021	0.768 ± 0.155	0.785 ± 0.111	*0.836 ± 0.135*	0.766 ± 0.093	0.780 ± 0.075	**0.869 ± 0.111**
misc.forsale	0.518 ± 0.025	0.552 ± 0.027	0.565 ± 0.146	*0.729 ± 0.102*	0.707 ± 0.119	0.706 ± 0.076	0.674 ± 0.072	**0.768 ± 0.128**
rec.autos	0.529 ± 0.033	0.720 ± 0.037	0.667 ± 0.166	0.775 ± 0.088	*0.812 ± 0.129*	0.762 ± 0.081	0.724 ± 0.091	**0.830 ± 0.121**
rec.motorcycles	0.506 ± 0.035	0.640 ± 0.028	0.802 ± 0.144	0.577 ± 0.102	*0.853 ± 0.119*	0.750 ± 0.097	0.814 ± 0.066	**0.868 ± 0.116**
rec.sport.baseball	0.517 ± 0.028	0.647 ± 0.031	0.779 ± 0.148	0.837 ± 0.078	*0.871 ± 0.113*	0.774 ± 0.080	0.790 ± 0.078	**0.887 ± 0.113**
rec.sport.hockey	0.513 ± 0.034	0.850 ± 0.025	0.823 ± 0.137	0.833 ± 0.096	0.918 ± 0.111	*0.936 ± 0.041*	0.932 ± 0.045	**0.947 ± 0.085**
sci.crypt	0.563 ± 0.036	0.696 ± 0.021	0.760 ± 0.146	0.768 ± 0.122	*0.808 ± 0.154*	0.804 ± 0.063	0.748 ± 0.088	**0.826 ± 0.142**
sci.electronics	0.506 ± 0.019	0.871 ± 0.017	0.555 ± 0.156	**0.940 ± 0.078**	*0.928 ± 0.090*	0.854 ± 0.053	0.828 ± 0.064	0.927 ± 0.088
sci.med	0.506 ± 0.019	0.621 ± 0.039	0.783 ± 0.125	0.832 ± 0.091	*0.862 ± 0.111*	0.772 ± 0.090	0.742 ± 0.101	**0.885 ± 0.107**
sci.space	0.547 ± 0.025	0.757 ± 0.034	0.818 ± 0.131	0.796 ± 0.110	0.820 ± 0.087	*0.888 ± 0.062*	**0.894 ± 0.063**	0.869 ± 0.112
soc.religion.christian	0.492 ± 0.034	0.590 ± 0.047	0.814 ± 0.138	*0.841 ± 0.140*	0.829 ± 0.122	0.726 ± 0.088	0.708 ± 0.100	**0.878 ± 0.113**
talk.politics.guns	0.477 ± 0.038	0.585 ± 0.060	0.747 ± 0.150	*0.806 ± 0.094*	0.782 ± 0.095	0.714 ± 0.074	0.708 ± 0.100	**0.814 ± 0.117**
talk.politics.mideast	0.559 ± 0.028	0.736 ± 0.026	0.793 ± 0.135	*0.830 ± 0.108*	0.825 ± 0.119	0.750 ± 0.084	0.784 ± 0.064	**0.867 ± 0.127**
talk.politics.misc	0.515 ± 0.037	0.704 ± 0.036	0.697 ± 0.152	0.720 ± 0.119	0.748 ± 0.140	0.788 ± 0.091	*0.806 ± 0.078*	**0.829 ± 0.135**
talk.religion.misc	0.554 ± 0.043	0.633 ± 0.035	0.739 ± 0.151	0.725 ± 0.104	*0.778 ± 0.114*	0.738 ± 0.074	0.746 ± 0.082	**0.821 ± 0.115**
average	0.515	0.679	0.726	0.791	*0.820*	0.767	0.766	**0.855**

It can be seen that BSN outperforms all competitors in all cases except for *sci.electronics*, where the accuracy of MInD (best) and MI-Net (second best) is higher by 1.3% and 0.1%, respectively. The average accuracy on all 20 datasets indicates that both MI-Net and BSN outperform the others.

Experiments

Table 4
Comparison of different methods (*mean* ± *std*) for bag classification on Messidor dataset.

mi-SVM	MI-SVM	miVLAD	miFV	MInD	MI-Net	Att.Net	Gated Att. Net	BSN
0.620 ±0.039	0.640 ±0.050	0.691 ±0.037	0.715 ±0.047	0.665 ±0.071	*0.730* ±0.051	0.703 ±0.041	0.698 ±0.048	**0.737** ±0.050

Table 4 shows that BSN achieves the best result on Medical image diagnosis

Table 5
Comparison of different methods (*mean* ± *std*) for bag classification on Colon Cancer dataset.

Method	Accuracy	precision	Recall	F-score
Instance+max	0.842 ± 0.021	0.866 ± 0.017	0.816 ± 0.031	0.839 ± 0.023
Embedding+max	0.824 ± 0.015	0.884 ± 0.014	0.753 ± 0.020	0.813 ± 0.017
Att. Net	**0.904 ± 0.011**	**0.953 ± 0.014**	*0.855 ± 0.017*	**0.901 ± 0.011**
Gated Att. Net	*0.898 ± 0.020*	*0.944 ± 0.016*	0.851 ± 0.035	*0.893 ± 0.022*
BSN	0.869 ± 0.008	0.820 ± 0.019	**0.983 ± 0.019**	0.886 ± 0.030

Table 5 presents Att. Net obtains the best precision (95.3±1.4%), whereas BSN obtains the best recall (98.3±1.9%).

Ablation study

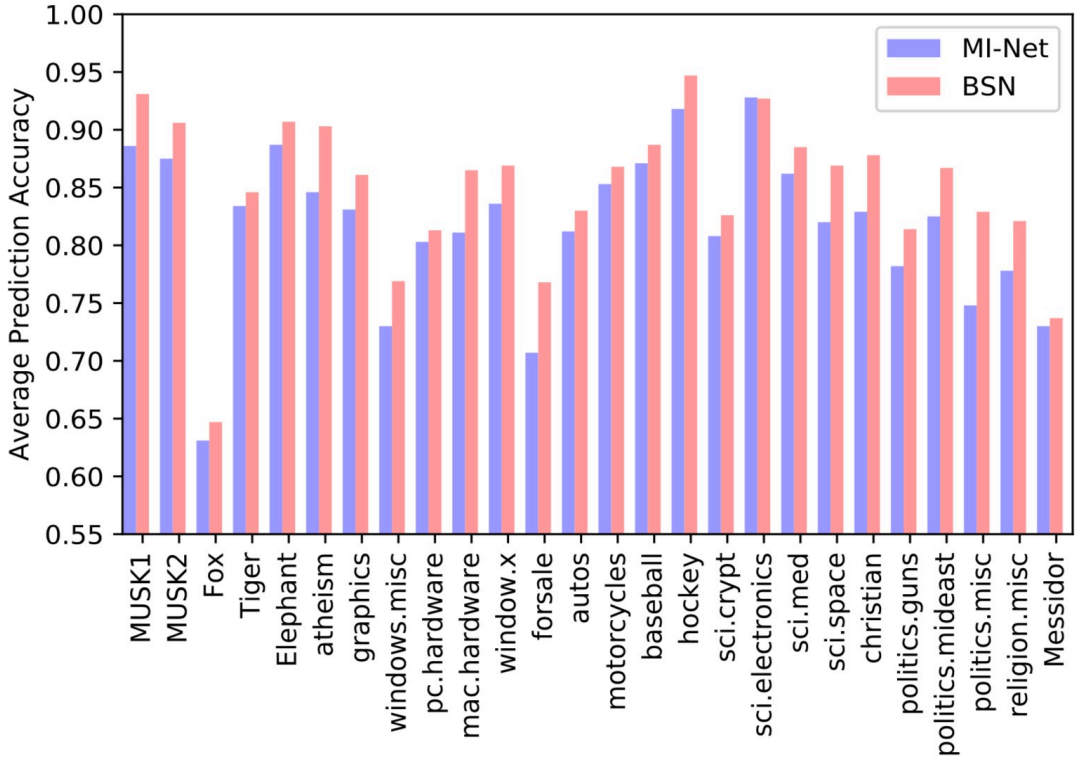

Fig. 3. Effectiveness of the proposed bag similarity representation on MUSK1, MUSK2, Fox, Tiger, Elephant, 20 Newsgroups, and Messidor datasets.

The proposed bag similarity representation are superior to other methods such as MI-Net To justify the argument and indicate the improvement, Fig. 3 offer a detailed comparison between the proposed BSN and MI-Net.

The average accuracy of BSN was improved. As both BSN and MI-Net are based on the same instance feature learning network, the results clearly demonstrate that bag similarity representation is useful for MIL problems.

Ablation study

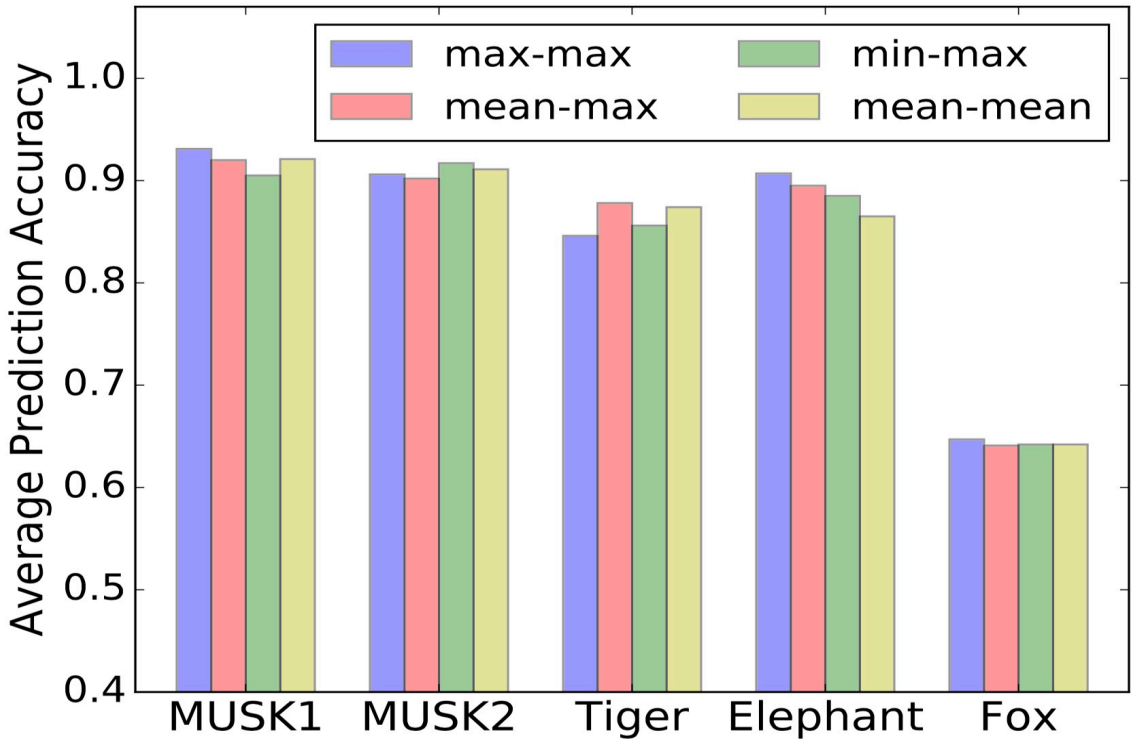

Fig. 4. Comparison of different MIL pooling methods on MUSK1, MUSK2, Fox, Tiger, and Elephant datasets.

The paper introduces four Hausdorff pooling methods to convert the bag similarity matrix into a bag-level feature vector: max–max, mean–max, min–max, and mean– mean pooling functions, which are all differentiable. Usually, the second operator in Hausdorff polling is preferably a max or mean function.

Bag similarity & Dissimilarity Representation for MIL

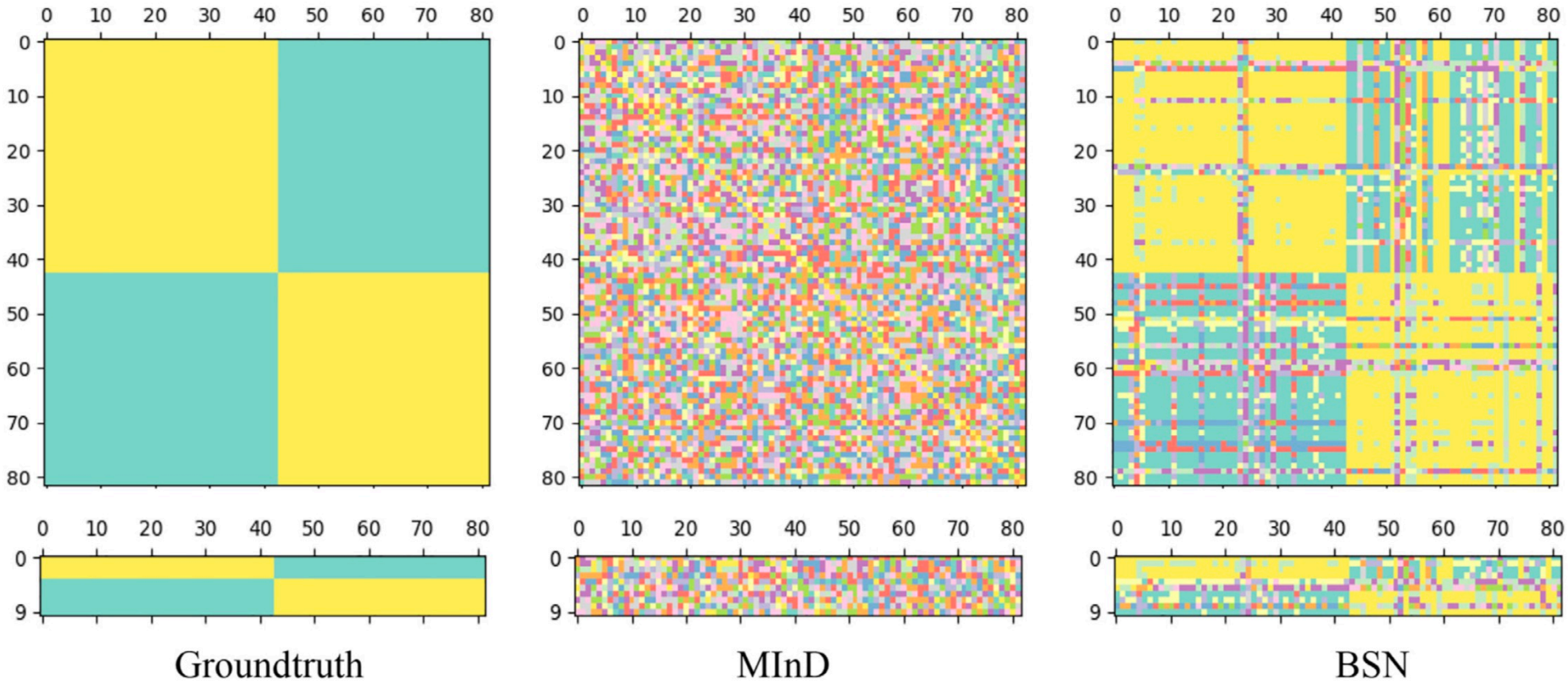

Fig. 1. Comparison between the fixed similarity metric (MInD) [4] and the learned similarity matrix by BSN. The **top** row shows the similarity matrices of training bags, and the **bottom** row shows the similarity matrices of testing bags. Even though the diagonal-blockness of the similarity matrix by BSN is not perfect, it is quite satisfactory, whereas MInD lacks this desired property.

Summary

- Proposed a learnable bag similarity and bag dissimilarities framework representation for MIL.;
- Point set distances and distribution distances are considered;
- Characteristics of MIL problems affect the effect of different algorithms.