Joint Clustering and Classification for Multiple Instance Learning

Karan Sikka
ksikka@ucsd.edu
Ritwik Giri
rgiri@ucsd.edu
Marian Bartlett
mbartlett@ucsd.edu

University of California San Diego
La Jolla, California
USA

Abstract

The Multiple Instance Learning (MIL) framework has been extensively used to solve weakly labeled visual classification problems, where each image or video is treated as a bag of instances. Instance Space based MIL algorithms construct a classifier by modifying standard classifiers by defining the probability that a bag is of the target class as the maximum over the probabilities that its instances are of the target class. Although they are the most commonly used MIL algorithms, they do not account for the possibility that the instances may have multiple intermediate concepts, and that these concepts may have unequal weighting in predicting the overall target class. On the other hand, Embedding-space (ES) based MIL approaches are able to tackle this issue by defining a set of concepts, and then embedding each bag into a concept space, followed by training a standard classifier in the embedding space. In previous ES based approaches, the concepts were discovered separately from the classifier, and thus were not optimized for the final classification task. Here we propose a novel algorithm to estimate concepts and classifier parameters by jointly optimizing a classification loss. This approach discovers a small set of discriminative concepts, which yield superior classification performance. The proposed algorithm is referred to as Joint Clustering Classification for MIL data (JC^{2}MIL) because the discovered concepts induce clusters of data instances. In comparison to previous approaches JC^{2}MIL obtains state-of-the-art results on several MIL datasets- Corel-2000, image annotation datasets (Elephant, Tiger and Fox), and UCSB Breast Cancer dataset.

1 Introduction

Many early approaches to visual classification were based on extracting a global descriptor, followed by learning a classification function using training labels [19, 20]. Since these methods describe an image/video as a whole *e.g.* GIST [19] and color histograms [20], they were often referred to as global representations. These approaches worked well for problems where the test object (or action) was visually uniform, but performed poorly when there were large variations in the object class due to factors such as occlusion, viewpoint changes or visual sub-categories [6, 11, 13]. A possible solution is to use local approaches where each image or video is represented by a set of localized visual descriptors [5, 6]. For example, as

Pages 71.1-71.12
DOI: https://dx.doi.org/10.5244/C.29.71

shown in Figure 2, a beach scene could be described by instances corresponding to a set of underlying regions. However, this setting differs from the standard supervised learning case where a unique class label is provided with each training instance. The only information available about the beach image is that at least one of the regions inside the beach scene corresponds to the target class. Since there is incomplete information regarding the class labels, this setting is often referred to as Weakly Supervised setting.

This setting is common in computer vision since complete human annotation is both costly and intensive, while labeling an image or video with a single global label is often more feasible. The data can be structured as a set of bags, each containing many instances, along with labels indicating the presence/absence of the object of interest in each bag. The learning problem in this case is referred to as Multiple Instance Learning (MIL) [1, 3]. Most of the previous MIL algorithms extended standard supervised learning algorithms to the MIL setting by assuming that the posterior probability of a bag containing a positive target class is a maximum over the probability of each of its instances. Thus if one of the instances has the target class with high probability, then the probability of the bag containing that target is also high irrespective of other instances. We shall refer to this as the *max* MIL assumption. This idea has been used to extend several supervised learning algorithms such as boosting [28] and logistic regression [8], to the MIL setting. Adapting the taxonomy proposed in [1], we shall refer to these algorithms as **Instance Space** (IS) based algorithms since they first define a classification function in the space of instances and extend it to the entire bag using the *max* assumption. IS-based MIL methods work well when the positive class can be described by a single target concept. However, the assumption of a single target concept might be too restrictive in several vision problems, e.g. a scene may comprise several local regions or an action may include multiple components. In such cases a bag is composed of heterogeneous concepts, which may contribute differently towards its classification. Consider the example of classifying a scene into beach or desert. Examples from both classes contain regions corresponding to sand and cloud, however in the case of the beach scene both water and sand must occur together. Moreover, these regions correspond to intermediate concepts that differ from the target label "beach".

To tackle the above issues, we focus on **Embedding Space** (ES) based MIL approaches [1] that embed each bag into a K-dimensional vector space. The procedure is illustrated in Figure 2, where a beach scene is represented as a bag of image regions. In this example each concept has an associated semantic meaning such as water, clouds or background. First, a similarity score between each instance and a concept is computed, which produces a **Concept-wise Instance Similarity** (CIS) for every instance. In the next step, the similarity between the bag and a concept is computed as the max CIS score, using the *max* MIL assumption similar to IS-based methods. The likelihood of each concept forms each dimension of an intermediate vector space, referred to as **concept space**. In other words, the set of concept likelihoods for the bag forms the embedding. After embedding each bag into the concept space, standard classifiers can be used to classify the overall target from the embedded representation. The concepts can take a number of forms. They could be cluster centers or dictionary atoms discovered by unsupervised methods such as k-means [5, 6, 9], or they could be concept prototypes learned by maximizing Diverse Density measure on MIL data [4, 17, 29]. The concepts induce a clustering of data instances into multiple categories as shown in Figure 2. This is related to dictionary learning based recognition approaches, where a dictionary is obtained using unsupervised algorithms. This dictionary is then used to encode features with the aim of improving recognition rate in the new space [6, 9].

In most previous ES-based approaches, the concepts and the classifier were obtained

independently. However, isolated learning of the classifier and the set of concepts (or dictionary) may not be optimal for the final classification task since the discovered concepts, which were optimized for a different objective (such as minimization of inter-cluster distance in k-means), may induce a concept embedding that is not suited to classification. [9, 16]. We propose to tackle this problem by introducing a novel ES-based MIL method that jointly learns the set of concepts and the classifier in a MIL setting. We refer to our algorithm as **Joint Clustering and Classification for Multiple Instance Learning** (JC²MIL). Our work makes the following contributions:

1. Proposes a framework to estimate concepts and classifier parameters by jointly optimizing a classification loss on the MIL data.

2. Shows that the current approach is able to yield state-of-the-art results on several MIL datasets by discovering discriminative concepts. The number of concepts are much smaller compared to the overcomplete set used in previous ES-based MIL algorithms [5].

To facilitate a fair comparison with the previous state-of-the-art ES-based MIL approaches [4, 5, 22], we use a RBF kernel based mapping function. It is also interesting to note that JC²MIL follows the line of recent work on task-driven dictionary learning [16].

2 Related Work

The idea behind IS-based methods is to construct an instance classifier by modifying a supervised learning algorithm using the MIL assumptions [1]. A number of algorithms have been proposed, such as MILBOOST [29], MI-logistic regression [8], MI-SVM [2], MI-Forests [14]), and used to tackle visual classification problems [23, 26, 27] . IS-based formulation has also been used to propose a mixture of linear [25, 26] or non-linear classifiers [17, 27] to solve the MIL problem. Although these algorithms were capable of detecting multiple concepts, they did not assume that different concepts can contribute differently towards the label of the bag.

On the other hand, the ES-based MIL algorithms are able to incorporate the above assumption by first embedding each bag into a concept space, followed by learning a standard classifier in this space [1]. A classic example of ES-based algorithms is the Bag of Words (BoW) model [6] that maps an image/video into a histogram using an unsupervised dictionary. Popular ES-based MIL approaches are based on the idea of extracting prototype(s) by maximizing the diverse-density (DD) function [17] on MIL data. The motivation is that a point with high DD is close to at least one instance inside a positive bag and far away from every instance in the negative bags. In this way, each prototype can also be identified as a positive concept. Chen *et al*. proposed DD-SVM [4] to discover several concepts through DD function and then used the corresponding concept space with SVM. An improvement was later proposed over DD-SVM called MILES [5], where the set of concepts included all the instances in the dataset, and the relevant instances were selected using sparse-SVM. A recent algorithm, called Dictionary based Multiple Instance Learning (DMIL) [22], learned a sparse reconstruction based dictionary by maximizing the DD function, and then used the sparse codes for each instance to embed a bag. Although this algorithm achieved excellent results, DMIL did not have an explicit notion of clustering owing to the use of a single mapping function for constructing the concept space.

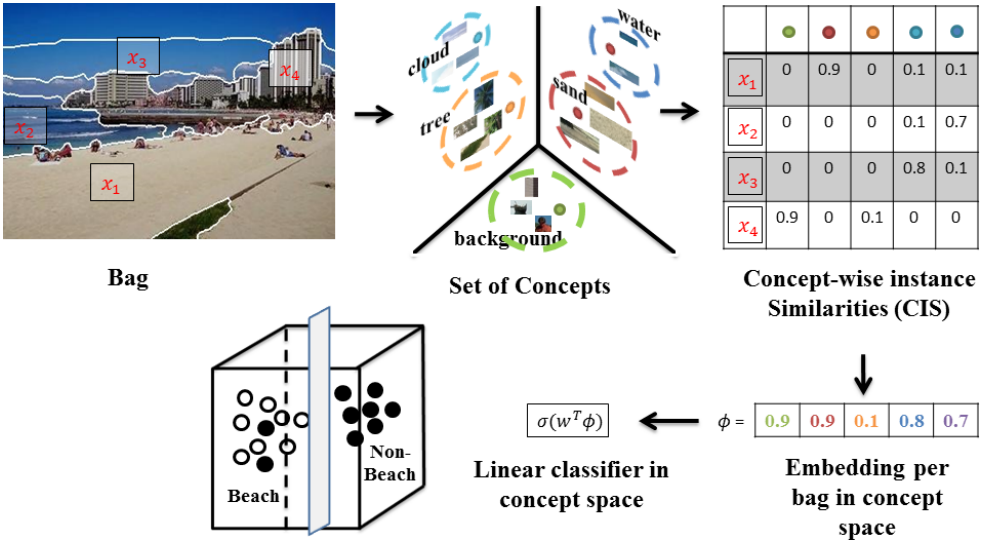

Figure 1: We illustrate the inherent idea in Embedding Space based MIL approaches. A beach scene, segmented into regions, is represented as a bag of instances, where each instance is the visual descriptor of the corresponding region. A set of concepts is then used to calculate a similarity between each instance and the concept, referred to as concept-wise instance similarity (CIS). The *max* MIL assumption is used to embed each bag into the concept space using the CIS. Classification can then be performed in the embedding space using standard classifiers (best viewed in color).

It is also worth mentioning about Bag Space (BS) based MIL algorithms [1] that calculate similarities between bags through kernels such as MI-kernel [10], mi-Graph [30], followed by the application of a Kernel-based learner such as SVM. These methods usually assume non i.i.d relationships between instances inside a bag and require the definition of a distance function. Although these methods have good performance, they are generally not able to do any instance classification. This paper focuses only on IS-based and ES-based MIL algorithms that allow both instance and bag classification.

Our work is also related to the recent work on task-driven dictionary learning [9, 16], where dictionary learning is coupled with the final task e.g. classification. In such a setting it is possible to achieve a good recognition rate without the use of an overcomplete dictionary (as in BoW [6] or MILES [5]) since the learned dictionary is already tuned to the final task. Similar extensions have been proposed in sparse-coding [16], BoW [15]. This work extends the previous ES-based MIL algorithms by discovering discriminative concepts that are learned simultaneously with the bag-level classifier.

3 Joint Clustering and Classification for Multiple Instance Learning(JC^{2}MIL)

3.1 Model

A MIL dataset is denoted as $B = \{X_i, y_i\}_{i=1}^{N}$, where X_i is the i^{th} bag, $y_i \in \{0,1\}$ is its binary label[1] and N is the cardinality of the dataset. Each bag X_i consists of N_i instances, denoted as $X_i = \{x_{ij}\}_{j=1}^{N_i}$, where $x_{ij} \in \mathcal{R}^d$ is a visual descriptor representing an instance. The target of this work is to jointly learn (1) a set of concepts, that are used to embed each bag into a concept space, and (2) a classifier that combines the embedding to produce a classification score for each bag. Our algorithm discovers discriminative concepts by learning them simultaneously with the classifier, that requires minimization of the classification loss on the training data.

The set of concepts is denoted as $\mathcal{C} = \{\mu_k\}_{k=1}^{K}$, consisting of K elements μ_k. The similarity between the k^{th} concept and instance x_{ij} is denoted as p_{ijk}. We purposely selected the RBF kernel, written as $p_{ijk} = \exp(-\frac{\gamma}{2}||x_{ij} - \mu_k||_2^2)$, since it was used in several previous ES-based methods- MILES [5], DD-SVM [4] and DMIL [22]). The CIS for the k^{th} concept are then used to derive the k^{th} embedding dimension, denoted as ϕ_{ik}, for the i^{th} bag using the *max* MIL assumption:

$$\phi_{ik} = \max_j p_{ijk} \tag{1}$$

The underlying idea is similar to the Instance based MIL algorithms where the probability of a bag is defined as the maximum over the probability of each of its instances. However, instead of using a single classifier or a concept, the probability is calculated with respect to multiple concepts. The vector containing the similarity score from all the K concepts for the i^{th} bag is denoted as $\phi_i = \{\phi_{ik}\}_{k=1}^{K} \in \mathcal{R}^K$, which also forms the embedding of the bag in the concept space. The classification score is then obtained by a linear classifier with parameter w. In this work we opted for a logistic regression classifier since it is able to provide good generalization by minimizing a differentiable loss function. The classification score for the i^{th} bag is converted to posterior probability, denoted as p_i, by using the sigmoid function as $p_i = \sigma(w^T \phi_i)$ [18].

3.2 Joint Optimization

We pose the problem as joint minimization of the classification loss with respect to the concepts and the classifier parameters. The classification loss includes the mean negative log-likelihood and a regularization term [18], and written as:

$$\mathcal{L}(B) = -\frac{1}{N}\sum_i (y_i \log p_i + (1-y_i)\log(1-p_i)) + \frac{\lambda}{2} w^T w \tag{2}$$

$$\{\mathcal{C}^*, w^*\} = \arg\min_{\mathcal{C},w} \mathcal{L}(B) \tag{3}$$

The above optimization problem is not jointly convex in both $\mathcal{C}$ and w, however it is convex in either while keeping the other variable fixed. Thus the minimization is performed via coordinate descent approach, where $\mathcal{L}$ is minimized alternatively with respect to both the

[1] Although the algorithm is formulated for binary classification problems, it can be extended to multiclass problems by learning one-vs-all binary classifiers.

variables. Since the original problem is non-convex, the final solution is dependent on the initialization. During our experiments we found that the initialization of the concepts with k-means (cluster centers) [7] was able to find relevant clusters and yield good results. The alternate minimization was performed using the BFGS algorithm, which is faster compared to gradient descent [21].

The gradient of $\mathcal{L}$ can be easily computed with respect to w (logistic regression [18]). The gradient of $\mathcal{L}$ with respect to each individual concept μ_k is expanded using the chain-rule of differentiation as:

$$\begin{aligned}\frac{\partial L}{\partial \mu_k} &= \frac{1}{N}\sum_i (p_i - y_i)\frac{\partial \phi_i}{\partial \mu_k} \\ &= \frac{1}{N}\sum_i (p_i - y_i) w_k \frac{\partial \phi_{ik}}{\partial \mu_k}\end{aligned} \tag{4}$$

The second expression follows since only the k^{th} dimension of ϕ_i is dependent on μ_k. ϕ_{ik} was computed in Equation 1 by taking a maximum over the per instance similarity scores (p_{ijk}) for the k^{th} concept. However, the non-differentiability of the maximum function poses a problem in estimating the above derivative. This problem was solved by using a soft-max approximation [3]. We used the Generalized Mean (GM) approximation instead of the frequently used NOR approximation since previously GM has been shown to provide better performance compared to NOR [3, 23]. As per the GM function, the embedding ϕ_{ik} is defined using p_{ijk} as $\phi_{ik} = (\frac{1}{N_i}\sum_j p_{ijk}^r)^{\frac{1}{r}}$, where r is a parameter controlling the degree of approximation. The inner-derivative in Equation. 4 is then written as:

$$\frac{\partial \phi_{ik}}{\partial \mu_k} = \frac{\phi_{ik}}{\sum_j p_{ijk}^r}\sum_j p_{ijk}^{r-1}\frac{\partial p_{ijk}}{\partial \mu_k} \tag{5}$$

The above derivative can be easily computed to write the derivative of the loss function with respect to k^{th} concept (μ_k) as:

$$\frac{\partial L}{\partial \mu_k} = -\frac{\gamma w_k}{N}\frac{\sum_i (p_i - t_i)\phi_{ik}}{\sum_j p_{ijk}^r}\sum_j p_{ijk}^r(\mu_k - x_{ij}) \tag{6}$$

Computing the above gradient is efficient since it can be written in terms of matrix products. For instance classification, we use the same procedure as introduced in MILES [5].

4 Experiments

To establish the empirical superiority of our proposed method over previous MIL algorithms, we tested it on five MIL datasets. The number of concepts and the RBF kernel parameter γ were tuned by using five fold cross-validation on the training set. The regularization and Generalized Mean parameter were set to a constant value ($\lambda = 10^{-5}$, $r = 15$). We found the algorithm to converge in about 20 iterations of coordinate descent. As a standard pre-processing step, the features were processed to have a zero-mean and unit variance. The concepts were initialized with 10 repetitions (for reproducing results) of k-means [7].

The first set of experiments were performed on the three image annotation datasets-Tiger, Fox and Elephant [2]. In these datasets, an image consists of a set of segments (or

blobs), each characterized by color, texture and shape descriptors. For an image lying in the positive class at least one of the blobs belongs to the target category. We used the same experimental protocol as described in [14], where average classification accuracy was reported using 5 randomly selected 10 fold cross-validation sets [2]. The comparative performance for these 3 datasets is shown in Table 1.

The second set of experiments were conducted on the Corel-2000 dataset that consists of 20 object categories with 100 images per category. In order to make a fair comparison, we used the same version of dataset as in [5], where each image was segmented into a number of regions, each of which were described by a 9 dimensional low-level feature vector consisting of color moments, gradient etc. Similar to [5], 5 random splits were generated by dividing the images into two equal parts, where (for each split) one part was used for testing and other for training. Since the present formulation only addresses binary classification problems, we conducted multiclass classification by training 20 per class one-vs-all binary classifiers. During test time, an image was classified as belonging to the category with the highest classification score. The results of different algorithms are shown in Table 2.

We also tested our algorithm on a public breast cancer dataset (UCSB Breast Cancer) with image samples taken from 32 benign and 26 malignant breast cancer patients [12]. Each image was divided into an equal-sized 7×7 patch and visual features such as SIFT, color histograms were extracted from each patch to form a 708 dimensional vector [12]. The dimensionality was reduced to 100 by application of Principal Component Analysis [7]. The task is to classify malignant and benign cancer images and the metric being reported is the mean area under the ROC curve (AROC) for a 10-fold stratified cross-validation (Table 3).

Method	type	Elephant	Fox	Tiger
mi-SVM [2]	IS	82	58	79
MI-SVM [2]	IS	81	59	84
MILBoost-NOR [28]	IS	73	58	56
EM-DD [29]	IS	78	56	72
MIForests [14]	IS	82	64	82
MILES [5]	ES	81	62	80
DMIL [22]	ES	**87**	68	**89**
JC^{2}MIL (Ours)	ES	**86***	**73**	**88***

Table 1: Comparison of our algorithm (% accuracy) with different Instance Space based (IS) or Embedding Space (ES) based algorithms on three image annotation datasets [2]. The second column shows the type of MIL algorithm (see Section 1).* The results were similar to DMIL in the case of the Elephant and the Tiger datasets since the standard deviation in both of these cases was around 1%.

5 Results and Discussion

Table 1 shows that our algorithm outperforms the performance of the best Instance Space based MIL algorithms on the MIL benchmark datasets by an absolute margin of 4% on Elephant, 9% on Fox and 4% on Tiger. Although DMIL seems to perform better than JC^{2}MIL on the Tiger and Elephant datasets by a margin of $\sim 1\%$, the results are statistically similar since both accuracies have a standard deviation of $\sim 1\%$. The performance improvement

[2]The random splits were downloaded from code provided by authors in [14].

Method	Corel-2000
MI-SVM [2]	54.6: [53.1 63.1]
MILES [5]	68.7: [67.3 70.1]
DD-SVM [4]	67.5: [66.1 68.9]
k-means-SVM [6]	52.3: [51.6 52.9]
DMIL [22]	70.2: [68.3 72.1]
JC²MIL (Ours)	**73.2**: [71.2 74.8]

Table 2: Evaluation (multiclass % accuracy) of Instance Space and Embedding Space based MIL algorithms on the Corel-2000 dataset [4] along with 95% confidence interval.

Method	UCSB Breast Cancer
MILBoost-NOR [23]	0.83
MI-SVM [2]	0.90
MILES [5]	0.74
JC²MIL (Ours)	**0.95**

Table 3: Evaluation (mean Area Under the ROC curve) on the UCSB Breast Cancer dataset [12].

on the Fox dataset is significant relative to DMIL, with a margin of 5%. This hike in performance could be explained by the presence of multiple target concepts in the Fox dataset, which is successfully captured by our algorithm. This argument is also verified by the classification accuracy on the Corel-2000 dataset (Table 2), since this dataset is known to have scene images with multiple target concepts, an example is shown in Figure 2. In this multiclass classification problem, our algorithm shows a clear performance improvement of 3% relative to the previous state-of-the-art results achieved by DMIL. This table also shows that the ES-based approaches perform much better compared to the IS-based MIL methods on the Corel-2000 dataset, *e.g.* MI-SVM achieves a performance of 54.6% compared to 73.2% by *JC²MIL* and 68.7% by MILES. This observation highlights our contention that the IS-based approaches are unable to effectively tackle problems with multiple concepts that may contribute unequally towards the classification of the bag.

Our algorithm also achieves the state-of-the-art AROC score of 0.95 on the UCSB Breast Cancer dataset. The results on this dataset are interesting since in all of the previous datasets (except Tiger) the performance of MILES was greater than or equal to MI-SVM and MILBOOST. This could be the case since MILES might be overfitting as a result of using a large number of concepts and unable to select relevant concepts using sparse-SVM. In this scenario our algorithm not only outperforms MILES but also perform better than the state-of-the-art IS-based algorithms.

5.1 Advantages of Discovering Discriminative Concepts

As discussed in Section 1, the primary advantage of discovering discriminative concepts through joint training is that the discovered concepts are already tuned to the final task leading to a performance improvement. This allows to achieve a good recognition rate by using a small number of concepts compared to the methods that separately learn the classifier and the concepts [5, 6, 22]. The overcompleteness is required in unsupervised dictionary based methods since it relaxes the classification problem by inducing a high dimensional embedding space where the data can be easily separated. On the other hand by incorporating label information during concept discovery, our algorithm is able to induce a low dimensional

embedding space that is highly discriminative. To quantitatively highlight this point, we compared the performance of JC^{2}MIL with a variant employing the same set of concepts, discovered via k-means, that were used to initialize our algorithm. A fair comparison was made by using the same algorithmic design (classifier and similarity kernel) as JC^{2}MIL. We shall refer to this variant as "k-means + LR (BoW)" since in principle it is similar to the Bag of Words model.

Figure 2 shows the performance of both the algorithms as a function of the dictionary size for the Corel-2000 dataset. This result supports our contention that the task-specific concepts discovered by our algorithm are able to yield a significant performance improvement over k-means + LR (BoW). Moreover, our algorithm obtains the state-of-the-art results (73.2%) with a small number of concepts (equal to 20) and the performance saturates for higher number of concepts. On the contrary the performance of the BoW variant increases with the size of concepts and reaches to a maximum of 70.1% (at a concept size of 3000), which is still 3% points below the performance of JC^{2}MIL. We have only shown results till a concept size of 50 for clarity of exposition.

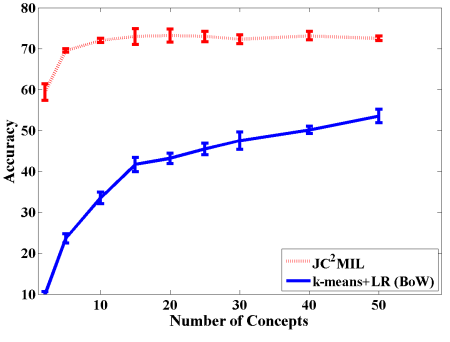

Figure 2: This graph highlights the advantage of learning discriminative concepts simultaneously with the classifier (see Section 5.1). The plot shows the performance of our algorithm and its variation using unsupervised concepts (k-means + LR (BoW)), on Corel-2000 dataset, as a function of the number of concepts. The performance of JC^{2}MIL reaches a maximum of 73.2% at a concept size of 20 and saturates for a higher number of concepts. On the other hand, the performance of the BoW variant reaches to a maximum of 70.1% at a concept size of 3000 and saturates thereafter.

6 Conclusion and Future Directions

This paper proposed a novel Embedding Space (ES) based Multiple Instance Learning (MIL) approach for visual classification problems. Unlike Instance-space (IS) based MIL approaches such as MILBOOST [28], the proposed method models the presence of multiple intermediate concepts that may contribute unequally towards predicting an object or an action. Arguing the advantages of learning concepts in a task-driven fashion, we proposed a novel approach for jointly learning the set of concepts and classifier parameters in a MIL setting. The proposed solution addresses an inherent issue in previous ES based methods where the concepts and the classifier were tuned independently, leading to concepts that may not be optimal for classification. We refer to our algorithm as Joint Clustering and Classification for Multiple Instance Learning (JC^{2}MIL) since the set of discovered concepts can be related to semantic clusters in the instance space. The performance advantages of JC^{2}MIL

were shown by reporting state-of-the-art results on several challenging MIL datasets. We further showed the advantages of discovering discriminating concepts in JC^{2}MIL compared to algorithms using unsupervised concepts.

We also observed that our algorithm outperformed unsupervised dictionary based ES methods by discovering a (relatively) small number of concepts. This allows our algorithm to be easily kernelized, making it possible to successfully cluster and classify fine-grained categories with high accuracy [9]. A possible research avenue that emerges with this work is towards using a more generic model for learning concepts (such as sparse coding based dictionaries [9, 16, 24]) in the proposed joint framework.

7 Acknowledgment

Support for this work was provided by NIH grant NIH R01NR013500. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the funding agency. We would like to thank Dr. Zhuowen Tu, Mohsen Malmir and Dr. Abhinav Dhall for helpful discussions. The authors in particular thank David Mateos-Núñez for helping with proofreading.

References

[1] Jaume Amores. Multiple instance classification: Review, taxonomy and comparative study. *Artificial Intelligence*, 201:81–105, 2013.

[2] Stuart Andrews, Ioannis Tsochantaridis, and Thomas Hofmann. Support vector machines for multiple-instance learning. In *Advances in neural information processing systems*, pages 561–568, 2002.

[3] Boris Babenko. Multiple instance learning: algorithms and applications, 2008.

[4] Yixin Chen and James Z Wang. Image categorization by learning and reasoning with regions. *The Journal of Machine Learning Research*, 5:913–939, 2004.

[5] Yixin Chen, Jinbo Bi, and James Ze Wang. Miles: Multiple-instance learning via embedded instance selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(12): 1931–1947, 2006.

[6] Gabriella Csurka, Christopher Dance, Lixin Fan, Jutta Willamowski, and Cédric Bray. Visual categorization with bags of keypoints. In *European Conference on Computer Vision Workshops*, volume 1, pages 1–2. Springer, 2004.

[7] Piotr Dollár. Piotr's Computer Vision Matlab Toolbox (PMT). http://vision.ucsd.edu/~pdollar/toolbox/doc/index.html.

[8] Zhouyu Fu and Antonio Robles-Kelly. An instance selection approach to multiple instance learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 911–918, 2009.

[9] Mehrdad J Gangeh, Ahmed K Farahat, Ali Ghodsi, and Mohamed S Kamel. Supervised dictionary learning and sparse representation-a review. *arXiv preprint arXiv:1502.05928*, 2015.

[10] Thomas Gärtner, Peter A Flach, Adam Kowalczyk, and Alex J Smola. Multi-instance kernels. In *International Conference on Machine Learning*, volume 2, pages 179–186, 2002.

[11] Kristen Grauman and Trevor Darrell. The pyramid match kernel: Discriminative classification with sets of image features. In *IEEE International Conference on Computer Vision*, volume 2, pages 1458–1465, 2005.

[12] Melih Kandemir, Chong Zhang, and Fred A Hamprecht. Empowering multiple instance histopathology cancer diagnosis by cell graphs. In *MICCAI*, 2014.

[13] Yan Ke and R. Sukthankar. Pca-sift: a more distinctive representation for local image descriptors. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages II–506–II–513 Vol.2, June 2004.

[14] Christian Leistner, Amir Saffari, and Horst Bischof. Miforests: multiple-instance learning with randomized trees. In *European Conference on Computer Vision*, pages 29–42. Springer, 2010.

[15] Xiao-Chen Lian, Zhiwei Li, Bao-Liang Lu, and Lei Zhang. Max-margin dictionary learning for multiclass image categorization. In *European Conference on Computer Vision*, pages 157–170. Springer, 2010.

[16] Julien Mairal, Francis Bach, and Jean Ponce. Task-driven dictionary learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(4):791–804, 2012.

[17] Oded Maron and Tomás Lozano-Pérez. A framework for multiple-instance learning. *Advances in neural information processing systems*, pages 570–576, 1998.

[18] Kevin P Murphy. *Machine Learning: a Probabilistic Perspective*. MIT press, 2012.

[19] Aude Oliva and Antonio Torralba. Building the gist of a scene: The role of global image features in recognition. *Progress in brain research*, 155:23–36, 2006.

[20] Otávio AB Penatti, Eduardo Valle, and Ricardo da S Torres. Comparative study of global color and texture descriptors for web image retrieval. *Journal of Visual Communication and Image Representation*, 23(2):359–380, 2012.

[21] Mark Schmidt. minfunc: unconstrained differentiable multivariate optimization in matlab). http://www.cs.ubc.ca/~schmidtm/Software/minFunc.html.

[22] A. Shrivastava, J.K. Pillai, V.M. Patel, and R. Chellappa. Dictionary-based multiple instance learning. In *IEEE International Conference on Image Processing*, pages 160–164, Oct 2014.

[23] Karan Sikka, Abhinav Dhall, and Marian Bartlett. Weakly supervised pain localization using multiple instance learning. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 1–8, 2013.

[24] Pablo Sprechmann and Guillermo Sapiro. Dictionary learning and sparse coding for unsupervised clustering. In *IEEE International Conference on Acoustics Speech and Signal Processing*, pages 2042–2045, 2010.

[25] Qifan Wang, Luo Si, and Dan Zhang. A discriminative data-dependent mixture-model approach for multiple instance learning in image classification. In *European Conference on Computer Vision*, pages 660–673. Springer, 2012.

[26] Xinggang Wang, Baoyuan Wang, Xiang Bai, Wenyu Liu, and Zhuowen Tu. Max-margin multiple-instance dictionary learning. In *International Conference on Machine Learning*, pages 846–854, 2013.

[27] Yan Xu, Jun-Yan Zhu, Eric Chang, and Zhuowen Tu. Multiple clustered instance learning for histopathology cancer image classification, segmentation and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 964–971, 2012.

[28] Cha Zhang, John C Platt, and Paul A Viola. Multiple instance boosting for object detection. In *Advances in neural information processing systems*, pages 1417–1424, 2005.

[29] Qi Zhang and Sally A Goldman. Em-dd: An improved multiple-instance learning technique. In *Advances in Neural Information Processing Systems*, pages 1073–1080, 2001.

[30] Zhi-Hua Zhou, Yu-Yin Sun, and Yu-Feng Li. Multi-instance learning by treating instances as non-iid samples. In *International Conference on Machine Learning*, pages 1249–1256, 2009.