2016 23rd International Conference on Pattern Recognition (ICPR)
Cancún Center, Cancún, México, December 4-8, 2016

Multiple Instance Learning with Multiple Positive and Negative Target Concepts

Andrew Karem and Hichem Frigui
Multimedia Research Lab, CECS Dept.
University of Louisville
Louisville, KY 400292
Email:adkare01@gmail.com or h.frigui@louisville.edu

Abstract*—We introduce a new algorithm that maps multiple instance data using both positive and negative target concepts into a data representation suitable for standard classification. Multiple instance data are characterized by bags which are in turn characterized by a variable number of feature vectors or instances. Each bag has a known positive or negative label, but the labels of any given instances within a bag is unknown. First, we use the Fuzzy Clustering of Multiple Instance data (FCMI) algorithm to identify K^+ positive target concepts, which represent points in the feature space that are close to instances from positive bags, and distant to instances from negative bags. We use a simple K-means clustering algorithm to identify K^- negative target concepts that supplement the positive target concepts. Next we demonstrate how the positive and negative target concepts can be used to embed each bag, which has a variable number of instances, into a feature vector with fixed dimension. A key advantage to embedded instance space feature vectors is that standard machine learning algorithms may be used in training and testing multiple instance data. Another advantage of our embedding is that it provides a simple and intuitive interpretation of the data. We show that using our feature embedding, coupled with standard classifiers such as support vector machines or k-nearest neighbors, can outperform state-of-the-art Multiple Instance Learning classifiers on benchmark datasets.

I. Introduction

In the Multiple Instance Learning (MIL) framework, an individual data sample is represented by a bag of feature vectors called instances. Each bag is associated with a given binary label (positive or negative), but labels of individual instances within a bag are unknown. This many-to-one relationship between feature vectors and data labels produces an inherent ambiguity in determining which instances in a given bag are responsible for its associated label.

Most of the existing MIL approaches rely on one of three models for discrimination: the instance space paradigm, the bag space paradigm, or the embedded instance space paradigm. Under the instance space paradigm, a decision boundary is learned in the feature space associated with instances (i.e., instance feature space) that provides optimal discrimination between positive and negative bags in the training set. For example, the Diverse Density algorithm [1] locates a point in the feature space, called a target concept, with strong correlation to instances from positive bags, and negative correlation to instances from negative bags. By contrast, the bag space paradigm assumes discriminative relationships between bags can be modeled using bag-to-bag mapping metrics. These metrics are used to form new feature vectors for each sample that remove MIL ambiguity and permit standard classification to be performed. A variation on the bag space paradigm is the embedded instance space paradigm, in which bags are not mapped to other bags, but instead to multiple feature vectors within the instance feature space (e.g. target concepts derived using the Diverse Density algorithm).

In this paper, we propose a new MIL classifier that is based on an efficient embedding of the instance feature space, using unsupervised learning methods to learn the most representative target concepts for embedding. Specifically, we use the Fuzzy Clustering of Multiple Instance (FCMI) algorithm [2] to learn a set of positive target concepts. Negative concepts are then learned by clustering instances from negative bags using a standard clustering algorithm such as K-Means [3] clustering. We show that a simple classifier, such as the K-Nearest Neighbor (k-NN), or support vector machine (SVM), when applied to the embedded space can provide results that are comparable to the state-of-the-art methods.

II. Related Work

MIL was formally introduced in [4] to predict bonding behavior of proteins in drug design problems. The authors proposed a classifier based on the Axis-Parallel Rectangles (APR) model. The APR seeks a bounding box in the instance feature space that includes at least one instance from each positive bag, but no or as few instances from negative bags as possible. APR is the first of many instance-space learning approaches [5], [6], which learn a decision boundary or region within the instance feature space that can discriminate between instances from positive bags and instances from negative bags. Another early instance space-based approach is the Diverse Density (DD) algorithm [1], which seeks points in the instance feature space with strong correlation to instances from positive bags and no or low correlation to instances from negative bags. These points, dubbed target concepts (TC), serve as loci for instance-level class labeling. Simple aggregation operations are then used to determine whether or not a given bag is positive or negative, depending on the labels of its instances. The most common of these are the most-likely cause estimate

978-1-5090-4847-2/16/$31.00 ©2016 IEEE

(MLE) metric [7], which assumes only the closest instance to a region or boundary within a bag defines its labels, and the NOISY-OR metric [1], [7], which models the label of a bag in a more probabilistic fashion with respect to its instance labels. The initial DD algorithm relied on a gradient descent approach to find a single optimal target concept. Several classifiers that extend or improve upon the DD approach have been proposed [5]. For example, in [8], the EM-DD utilizes an expectation maximization approach to find the target concept in a more efficient manner.

The aforementioned instance space-based algorithms rely on the standard multiple instance (SMI) assumption [5], which states that a positive bag must contain at least one positive instance [5]; the labels of remaining instances are irrelevant. This representation is not ideal when multiple instances in a positive bag define its label [5], [9], [10], for example when treating facial recognition [7], [11] or object detection [12], [13] as MIL applications. Other instance space-based methods abandon the SMI assumption and instead adopt a collective multiple instance (CMI) assumption in linking instance labels to bag labels. Under the CMI assumption, each instance contributes to its bags label. This contribution can be as simple as a mean aggregation (i.e. each instance plays an equal role in determining the bags label) [10], or be based on other characteristics of the data such as, for example, the number of other instances in a bag, number of instances in the dataset, or proportion of bags in the dataset that are positive [14], [15]. Unfortunately, the ideal weighting scheme may be drastically different from one dataset to the next [5], [15]. Furthermore, the number of instances that contribute to a bags label may vary even within a dataset. Instance space-based methods, whether following the SMI assumption or CMI assumption, are inadequate in addressing such data.

An alternative to the instance space paradigm that has emerged in MIL is the bag space paradigm [5], [6], [9]. In this paradigm, each bag is mapped to an N-dimensional feature vector based on a bag-to-bag comparator metric with respect to all N bags within the training data. In the bag space paradigm, decisions with regards to class discrimination are made at the bag level, using standard classifiers. The comparator metric for two bags is typically defined as a function of the distances between their individual instances. For example, the Citation k-NN classifier [16], uses the Hausdorff distance for mapping and a modified k-NN that utilizes both references and citers to derive a class label. Another bag-space approach is the Multiple Instance Dissimilarity (MInD) [9], which is a framework in which bags are treated as one of either an individual point-set, a distribution of instances, or an attributed graph. The authors in [9] utilize distance-based dissimilarities in bag-to-bag comparisons to bypass information loss imposed on outlying data by typical RBF similarity metrics. A key advantage to the bag space paradigm is that the mapped bag representation removes the instance-level ambiguity from the problem. This permits standard learning approaches (e.g. kNN, SVM, etc.) to be utilized on the transformed data. Despite these advantages, the bag space paradigm has serious drawbacks. First, the mapping is lossy at the instance space level. For example, in the case of the minimum Hausdorff distance, considering only the two closest instances from two bags may ignore useful information from secondary instances [9]. Moreover, bag-space transformations may cause a loss of solution interpretability, since we cannot necessarily explain a bag-space decision boundary in terms of the instance feature space.

An alternative to bag-space and instance-space paradigms is the embedded instance space paradigm [5], [9]. Similar to a bag-space approach, an embedded instance space approach maps each bag to a single feature vector. The difference is that the embedding utilizes target concepts in the instance space for this remapping, rather than bags. The selection of appropriate TCs is a primary factor in determining classification accuracy for embedded approaches. One of the earliest embedded approaches, DD-SVM [17], locates candidate TCs across multiple individual runs of the DD algorithm with distinct starting points, and consolidates the unique points derived by this process into a set of optimal TCs. Unfortunately, DD optimization is both sensitive to noise and prone to local minima. When 2 distinct TCs are present, DD has a high probability to converge to the mid-point between the two instead of to the one that is closer to the initial point [2]. This implies that restarting the DD algorithm with multiple starting points may fail to locate optimal TCs for any given MIL data.

To overcome the sensitivity of DD and learn the TCs when the MIL data is noisy and multi-modal, the authors in [18] use the K-means [3] algorithm to cluster all instances from positive bags into k clusters and assume that the centers of the largest cluster correspond to the TC. Unfortunately, for most applications, this assumption is not valid as large clusters may correspond to clusters of negative instances (within the positive bags), and positive instances may be split over multiple smaller clusters. Another embedded instance space MIL implementation is MILES [7], which considers each instance from both positive and negative bags as a potential TC, and computes an embedded similarity between the bags in the dataset and each potential TC. A sparse SVM is then utilized to weight and select which instances make optimal target concepts. MILES accommodates both positive and negative target concepts, and permits bags to possess membership across multiple target concepts. Despite these advantages, MILES may not be practical when bags have a very large number of instances, and as defined in [7], relies on a constant scaling vector for computing the embedded similarities.

III. Multiple Instance Learning with Multiplpe Positive and Negative Target Concepts

In this paper, we propose an embedded instance space algorithm that uses multiple positive and negative TCs, and addresses some of the limitations of existing embedded instance space approaches. Negative concepts are learned by clustering instances from negative bags using standard clustering algorithms such as the K-Means [3] . Positive concepts are learned from both positive and negative bags using a multi-concept diverse density algorithm [2].

Let $\mathcal{B} = \{B_1, \cdots, B_n, \cdots, B_N\}$ be the set of all bags in the dataset. Each bag $B_n = \{b_{n1}, \cdots, b_{ni}, \cdots, b_{nI}\}$ consists of I instances[1] and each instance is an F-dimensional feature vector. A bag is positive, denoted B_n^+, if at least one of its instances is positive; a bag is labeled negative, denoted by B_n^-, if none of its instances are positive. We assume that the data consist of N_{pos} positive and N_{neg} negative bags. Hence, our data may be represented in full as the union of $\mathcal{B}^+ = \{B_1^+, \cdots, B_{N_{pos}}^+\}$ positive bags and $\mathcal{B}^- = \{B_1^-, \cdots, B_{N_{neg}}^-\}$ negative bags.

A. Learning Positive Target Concepts

As outlined above, a positive target concept (TC) is point in the feature space that is densely populated by at least one instance from as many positive bags as possible, and as few instances from negative bags as possible. We use the Fuzzy Clustering of Multiple Instance (FCMI) algorithm [2] to learn multiple TCs that collectively maximize a multi-concept diverse density metric. Unlike approaches that attempt to locate distinct target concepts with multiple initializations (e.g. DD-SVM [7]), the FCMI discovers multiple TCs simultaneously.

FCMI assumes that the data can be described by K target concepts $\mathcal{T}^+ = \{t_1, \cdots, t_k, \cdots, t_K\}$. Each bag, B_n, is assumed to have a fuzzy membership u_{kn} representing the degree to which it is associated with concept t_k.

Let $\mathbf{U}$=$[u_{kn}]$ for $k = 1, \cdots, K$ and $n = 1, \cdots, N$. FCMI seeks to maximize the multi-target concept Diverse Density objective function defned by

$$MDD(\mathbf{T}^+, \mathbf{U}) = \prod_{n=1}^{N} \prod_{k=1}^{K} (Pr(t_k|B_n))^{u_{kn}^m}. \qquad (1)$$

subject to the constraints

$$u_{kn} \in [0, 1], \quad \text{and} \quad \sum_{k=1}^{K} u_{kn} = 1. \qquad (2)$$

In (1), $m \in (1, \infty)$ is a parameter that controls the fuzziness of the partition [19]. Using the NOISY-OR model [20], we let

$$Pr(t_k|B_n) = \begin{cases} 1 - \prod_{i=1}^{I}(1 - Pr(b_{ni} \in t_k)) & if\, B_n \in B^+ \\ \prod_{i=1}^{I}(1 - Pr(b_{ni} \in t_k)) & if\, B_n \in B^- \end{cases} \qquad (3)$$

In (3), $Pr(b_{ni} \in t_k)$ can be computed based on the similarity between instance b_{ni} and target concept t_k. If we let each t_k be characterized by a representative feature vector (e.g. centroid), c_k and a scaling vector s_k, then

$$Pr(b_{ni} \in t_k) = e^{-\left(\sum_{j=1}^{F} s_{kj}(b_{nij} - c_{kj})^2\right)} \qquad (4)$$

To avoid computation involving extremely small values in maximizing (1), we instead minimize its negative log-likelihood:

$$\begin{aligned} J(\mathbf{T}^+, \mathbf{U}) &= -log(MDD(\mathbf{T}, \mathbf{U})) \\ &= \sum_{n=1}^{N} \sum_{k=1}^{K} u_{kn}^m \{-log(Pr(t_k|B_n))\} \end{aligned} \qquad (5)$$

subject to (2).

There is no closed form solution for the minimization of (5). Instead, FCMI uses an iterative approach that alternates between optimization of the fuzzy membership matrix U, given a static set of target concepts T, and the optimization of the target concepts in T for a fixed U. A more detailed description of the FCMI algorithm can be found in [2]. After convergence, the FCMI identifies K TCs. Each concept t_k is characterized by a center c_k and scaling vector s_k. For the remainder of this paper, we use K^+ to refer to the K target concepts identified by FCMI.

B. Learning Negative Target Concepts

In many MIL applications, positive bags can be characterized by one or multiple positive TCs. In this case, new test bags can be labeled based on their proximity to TCs learned from training data as outlined in the previous section. However, in some MIL applications, negative TCs can provide equally important information by virtue of their dissimilarity to the true positive instances within positive bags. A negative target concept can then be defined as a point in the feature space that is as similar to as many instances from negative as possible, but not necessarily by its proximity to instances from positive bags, as many of these are assumed to be "background" instances. [2]

In [9], the authors identify three types of MIL concept datasets, and three corresponding behaviors with respect to positive and negative TCs. The first of these, "Concept" dataset, is one which satisfies traditional MIL assumptions, where a single or few instances from each positive bag determines its label, and the rest of the instances are "background." For such cases, traditional instance similarity functions, such as the RBF kernel utilized in [7], and a most-likely estimate function to each positive target concept (i.e. one in which only the nearest instance to each target concept is considered) are sufficient. The second dataset is "distribution," for which multiple instances from positive bags are proximal to the ideal TCs in comparison to instances from negative bags. In such cases, the most-likely estimate function is insufficient to describe the bag similarities to TCs, and proximity of individual instances to negative TCs may be a useful metric for discrimination, since instances from positive bags are less likely to occur in areas densely populated by instances from negative bags. The last dataset class is multi-concept, which has multiple positive instances substantially away from regions

[1] For simplicity in illustration, we assume each bag has the same number of instances. This need not be the case, and seldom is in practice.

[2] As an alternative, negative target concepts may be defined by their lack of proximity to at least one instance from as many positive bags as possible, but incorporating this definition is left as future work.

populated densely by instances from positive or negative bags, often regarded as data outliers. In [9], the authors argue that comparisons to other positive concepts for such data samples are insufficient, and instead they are better represented by their outlying dissimilarity to regions densely populated by negative instances (e.g. in our formulation, negative TCs.)

To address the above cases, we use the training data to learn supplemental negative TCs. Identification of negative TCs is more straightforward than that of positive TCs, as negative bags contain only negative instances. Thus, we simply apply the K-means clustering algorithm [3] to cluster all instances within the negative bags into K^- clusters.

C. Embedded Feature Space

Target concepts can be used to transform the bag-of-instances representation of MIL datasets into a set of standard feature vectors suitable for use with conventional classifiers. Assuming K^+ positive target concepts and K^- negative target concepts, each bag will be remapped to a feature vector of fixed size. Either similarity (e.g. probability) or dissimilarity (e.g. distance) can be used to represent the relationship between a target concept and a bag. In this paper, we use distance-based mapping because similarity-based metrics can result in information loss as all dissimilar bags (with significantly different distance values) map to zero.

One intuitive distance mapping for embedding features follows the most-likely estimate approach:

$$dmin(B_n, t_k) = \min_{i=1}^{I} d(b_{ni}, t_k)\,,\, for\, t_k \in \{T^+ \cup T^-\} \tag{6}$$

That is, each bag is mapped to a feature vector with (K^+ + K^-) dimensions based on the minimum distance across all of the instances to each TC.

We note that the distance in (6) applies to both positive ($t_k \in T^+$) and negative ($t_k \in T^+$) target concepts. In the former case, the distance provides information relevant to the ”concept” dataset paradigm outlined earlier, in which a single representative instance from each bag should be close to at least one positive TC. The latter case attempts to address the ”distribution” dataset paradigm, as we expect negative bags to have more instances close to distribution-based background or negative TCs. In addition to (6), we also include one aggregation-based feature for positive TCs:

$$dmin(B_n, \mathbf{T}^+) = \min_{k=1}^{K^+} \min_{i=1}^{I} d(b_{ni}, t_k) \tag{7}$$

That is, each bag will be mapped to an additional feature based on the smallest distance between any of its instances and any of the K^+ positive TCs. The distance in (7) can be regarded as a variation of the Hausdorff set distance applied to a set of target concepts as opposed to a set of instances in a bag. If at least one distance is small to at least one positive target concept, it is a strong indication that the bag is positive regardless of its distance to the remaining TCs.

To compute the distance $d(b_{ni}, t_k)$ between instance b_{ni} and target concept t_k in (6) and (7), we rely on a scaled Euclidean distance function:

$$d(b_{ni}, t_k) = \sum_{j=1}^{F} s_{kj}(b_{nij} - c_{kj})^2 \tag{8}$$

where $j = 1..F$ are the individual dimensions in the instance feature space. For positive TCs (i.e. $t_k \in \mathcal{T}^+$) we use the c_k and s_k derived from FCMI optimization for each target concept. For negative target concepts (i.e. $t_k \in \mathcal{T}^-$), we use the cluster centroid for c_k and the standard deviation across the features for each cluster using all instances assigned to it as an estimate of the cluster's scale s_k.

Given K^+ positive target concepts and K^- negative target concepts, combining (6) and (7) provides us with a $(K^+ + K^- + 1)$-dimensional feature vector for each bag. Any standard classification algorithm can then be used to classify the embedded MIL data.

IV. Experimental Results

We processed our embedded feature approach on five benchmark MIL datasets. The first two of these, MUSK-1 and MUSK-2 [4], [21], address the drug activity prediction application [4]. Each bag in MUSK-1 and MUSK-2 represents either a molecule with a musky scent (positive), or non-musky molecule (negative). Each molecule is described by multiple shape metrics describing low-energy conformations (instances), but it is unknown for positive molecules which individual conformation is responsible for the musky scent. The remaining three datasets, FOX, TIGER, and ELEPHANT, represent three segmented COREL image datasets [21], [22]. Each image is represented by a bag and each segmented image region is characterized by a 230-dimensional feature vectorand treated as an instance. Table I summarizes characteristics for these datasets.

TABLE I
SUMMARY STATISTICS OF BENCHMARK DATASETS

	MUSK-1	**MUSK-2**	**FOX**	**ELEPHANT**	**TIGER**
# Pos. Bags	47	39	100	100	100
# Neg. Bags	45	63	100	100	100
Avg. # Inst Per Bag	5.17	64.69	6.60	6.96	6.10

To validate our proposed embedding, we applied three standard classifiers to the mapped feature space. The first of these was a simple k-nearest neighbors classifier, the second was a modified citation-KNN based approach, which takes into account both references and citers [16] in assigning a label to a bag, and the third was a RBF-kernel based support vector machine. In our results, we refer to these Multiple Target Concept Embedding (MTC) approaches as MTC-KNN, MTC-CKNN, and MTC-SVM respectively. We compare our results

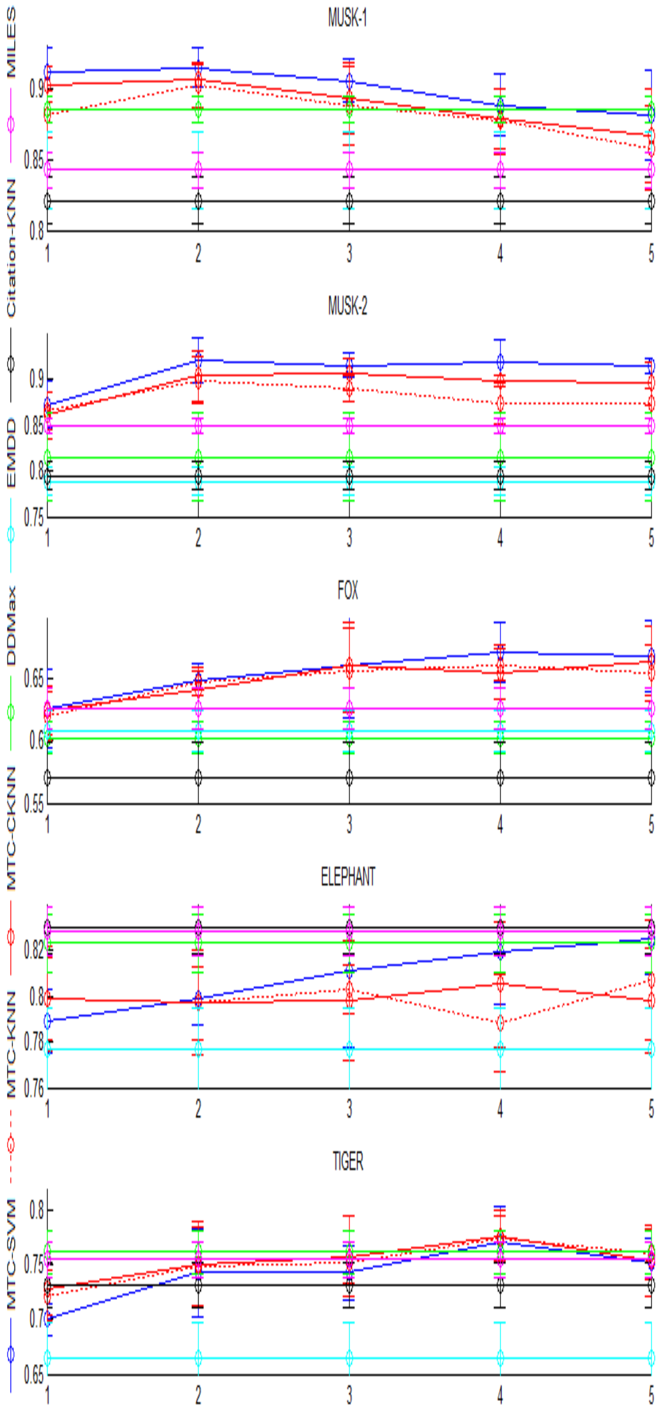

Fig. 1. Algorithm performance on benchmark datasets.

directly to those obtained through execution of other state-of-the art MIL algorithms. These are the DDMax [1], EM-DD [8], Citation KNN MIL [16] and MILES [7] classifiers. For all of these algorithms, we used the implementations available at [23]. For each dataset, we applied a 10-fold cross-validation scheme for training and testing. We repeated this process 5 times using random partitioning of the 10 folds, and we report the average accuracy and standard deviation for the 5 runs. For MTC-SVM parameterization, we used a simple RBF kernel with $\sigma = 1$. For MTC-KNN and MTC-CKNN, we use K=15 nearest neighbors. We note that we tried multiple values for K ranging from 10 to 20 and the results were comparable. With regards the FCMI TC selection, we varied K^+ from 1 to 5, and recorded the results using each value for each dataset and partition. For the negative TCs, we fixed $K^- = 5$, but note that our results showed similar and stable performance for as few as $K^- = 3$ or as many as $K^- = 7$ negative target concepts. We also compare our results to the aforementioned state-of-the-art methods. For all algorithms, we report the accuracy average over 5 runs with random cross-validation partitioning, as well as the standard deviation.

Figure 1 compares the classification accuracy of the proposed mapping as we vary K^+ from 1 to 5. We also compare our results to the aforementioned state-of-the-art methods. For all algorithms, we report the accuracy average over 5 runs with random cross-validation partitioning, as well as the standard deviation.

One observation to make regarding these results is that MTC-SVM, MTC-KNN, and MTC-CKNN outperform the existing algorithms significantly for the two drug-design datasets (MUSK-1 and MUSK-2), and one of the COREL datasets (FOX). For the TIGER dataset, MTC-SVM and MTC-CKNN outperform the other algorithms at $K^+ = 4$ TCs. For the ELEPHANT dataset, only the MTC-SVM is competitive at $K^+ = 4$ or $K^+ = 5$ TCs. Another observation to make is that MTC-CKNN is consistently better than MTC-KNN for each value for K^+ and each dataset. We also note that the optimal number of positive target concepts varies according to the dataset being processed. For example, for the MUSK data, $K^+ = 2$ generally provides peak accuracy, whereas for the three COREL datasets, $K^+ = 4$ provides better performance. These results are intuitive since the image categories require multiple target concepts to account for the large intra-class variations within their visual features. Table II summarizes our results when K^+ is fixed to 2 for the MUSK data and 4 for the COREL data.

TABLE II
COMPARISON OF THE CLASSIFICATION ACCURACY FOR OUR PROPOSED EMBEDDING USING 3 DIFFERENT STANDARD CLASSIFIERS AND 4 OTHER MIL CLASSIFIERS

	MUSK-1	MUSK-2	FOX	ELEPHANT	TIGER
MTC-SVM	**91.54%**	**92.09%**	**67.20%**	81.90%	77.10%
MTC-KNN	90.30%	89.88%	66.20%	78.80%	77.50%
MTC-CKNN	90.65%	90.32%	65.60%	80.50%	**77.70%**
DDMax	88.61%	81.53%	60.30%	82.30%	76.20%
EMDD	84.35%	78.82%	60.80%	77.70%	66.50%
Citation-KNN	82.17%	79.41%	57.10%	**83.00%**	73.20%
MILES	84.35%	84.90%	62.70%	82.80%	75.50%

As can be seen, FCMI-SVM with positive and negative target concepts shows substantial performance improvement for three datasets (MUSK-1, MUSK-2, and FOX), is as good or slightly better for one dataset (TIGER) and only slightly worse for the last (ELEPHANT).

V. Conclusions

We proposed an approach to multiple instance classification that utilizes both positive and negative target concepts. First, we described how the FCMI algorithm could be used to synthesize a set of K^+ positive target concepts from a MIL dataset, and how K-means clustering could used to synthesize a set of K^- negative target concepts from the same MIL data. Then, we showed how simple distance and aggregation operators could used to map each sample in the dataset to a $(K^+ + K^- + 1)$-dimensional feature vectors. These remapped feature vectors allow for standard classification approaches, such as the SVM or k-nearest Neighbors, to be applied to MIL data.

Using five benchmark datasets, we demonstrated that our approach is competitive with similar existing MIL approaches and outperforms them on four of the five datasets. One key advantage of our proposed embedding approach is that it provides interpretation of the data and justification of the classification results.

In our current approach, we vary the number of positive target concepts (K^+) and select the optimal value. Future work will include a target concept selection approach aimed at automating the ideal value for K^+. This could be achieved by treating K^+ as a variable and optimizing it within the FMCI algorithm. We will also explore alternative ways to embed features using the derived positive and negative target concepts.

Acknowledgment

This work was supported in part by U.S. Army Research Office Grants Number W911NF-13-1-0066 and W911NF-14-1-0589. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office, or the U.S. Government.

References

[1] O. Maron and T. Lozano-Pérez, "A framework for multiple-instance learning," *Advances in Neural Information Processing Systems*, vol. 10, no. 1, pp. 570–576, 1998.

[2] A. Karem and H. Frigui, "Fuzzy clustering of multiple instance data," in *Fuzzy Systems (FUZZ-IEEE), 2015 IEEE International Conference on*. IEEE, 2015, pp. 1–7.

[3] J. MacQueen, "Some methods for classification and analysis of multivariate observations," Berkeley, Calif., pp. 281–297, 1967.

[4] T. Dietterich, R. Lathrop, and T. Lozano-Prez, "Solving the multiple instance problem with axis-parallel rectangles," *Artificial Intelligence*, vol. 89, pp. 31–71, 1997.

[5] J. Amores, "Multiple instance classification: Review, taxonomy, and comparative study," *Artificial Intelligence*, vol. 201, pp. 81–105, 2013.

[6] J. Foulds and E. Frank, "A review of multi-instance learning assumptions," *The Knowledge Engineering Review*, vol. 25, no. 01, pp. 1–25, 2010.

[7] Y. Chen, J. Bi, and J. Z. Wang, "Miles: Multiple-instance learning via embedded instance selection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 12, pp. 1931–1947, 2006.

[8] Q. Zhang and S. A. Goldman, "Em-dd: An improved multiple-instance learning technique," in *In Advances in Neural Information Processing Systems*. MIT Press, 2001, pp. 1073–1080.

[9] V. Cheplygina, D. M. Tax, and M. Loog, "Multiple instance learning with bag dissimilarities," *Pattern Recognition*, vol. 48, no. 1, pp. 264–275, 2015.

[10] R. C. Bunescu and R. J. Mooney, "Multiple instance learning for sparse positive bags," in *Proceedings of the 24th international conference on Machine learning*. ACM, 2007, pp. 105–112.

[11] P. Wohlhart, M. Kostinger, P. Roth, and H. Bischof, "Multiple instance boosting for face recognition in videos," in *Pattern Recognition*, ser. Lecture Notes in Computer Science, R. Mester and M. Felsberg, Eds. Springer Berlin Heidelberg, 2011, vol. 6835, pp. 132–141.

[12] C. Zhang, J. C. Platt, and P. A. Viola, "Multiple instance boosting for object detection," in *Advances in neural information processing systems*, 2005, pp. 1417–1424.

[13] B. Babenko, M.-H. Yang, and S. Belongie, "Visual tracking with online multiple instance learning," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 983–990.

[14] E. Frank and X. Xu, "Applying propositional learning algorithms to multi-instance data," 2003.

[15] J. R. Foulds, "Learning instance weights in multi-instance learning," 2008.

[16] J. Wang and J.-D. Zucker, "Solving multiple-instance problem: A lazy learning approach," 2000.

[17] Y. Chen, J. Z. Wang, and D. Geman, "Image categorization by learning and reasoning with regions," *Journal of Machine Learning Research*, vol. 5, pp. 913–939, 2004.

[18] Z. Li, G.-H. Geng, J. Feng, J.-y. Peng, C. Wen, and J.-l. Liang, "Multiple instance learning based on positive instance selection and bag structure construction," *Pattern Recognition Letters*, vol. 40, pp. 19–26, 2014.

[19] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*. Norwell, MA, USA: Kluwer Academic Publishers, 1981.

[20] S. Srinivas, "A generalization of the noisy-or model," *Proceedings of the Ninth Conference on Uncertainty in Artificial Intelligence*, pp. 208–218, 1993.

[21] "Benchmark mil datasets," http://www.miproblems.org/general/musk-fox-tiger-and-elephant/, accessed: 2016-04-14.

[22] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support vector machines for multiple-instance learning," in *Advances in neural information processing systems*, 2002, pp. 561–568.

[23] D. Tax and V. Cheplygina, "MIL a Matlab toolbox for multiple instance learning," Jun 2015, version 1.1.0. [Online]. Available: http://prlab.tudelft.nl/david-tax/mil.html