ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

A New Approach for Multiple Instance Learning based on a Homogeneity Bag Operator

A. W. C. Faria[1], D. Menotti[2], A. P. Lemos[1] and A. P. Braga[1]

1- Graduate Program in Electrical Engineering - Federal University of Minas Gerais
Av. Antônio Carlos 6627, 31270-901 - Belo Horizonte, MG, Brazil

2- Computing Department - Federal University of Ouro Preto
Campus Universitário, 35400-000 - Ouro Preto, MG, Brazil

Abstract. Multiple Instance Learning (MIL) proposes a new paradigm when instance labeling, in the learning step, is not possible or infeasible, by assigning a single label (positive or negative) to a set of instances called *bag*. In this paper, an operator based on homogeneity of positive bags for MIL is introduced. Our method consists in removing instances from the positives bags according to their similarity with the ones from the negative bags. The experimental results show that our operator always increases the accuracy of the Citation kNN algorithm achieving the best results in 2 out of 4 datasets when compared with other classic methods in the literature.

1 Introduction

In Supervised Learning (SL) every sample $\mathbf{x}_i$ in the training set $D_L = \{\mathbf{x}_i, y_i\}_{i=1}^N$ is tagged with a class label y_i, so that there is a one-to-one relation between input vectors and class labels. In Multiple Instance Learning (MIL) [1], *bags* of samples are labeled instead of individual samples, what represents an additional challenge for current learning algorithms, since most of them rely on $\mathbf{x}_i \rightarrow y_i$ relations in order to induce mapping functions.

Positive bags B_i^+ contain at least one positive sample and negative bags B_i^- contain only negative samples, so the negative samples within a negative bag are labeled, whereas the samples within a positive bag are not labeled. This relatively new learning paradigm has been described first by Dietterich *et al.* [1] in the context of a drug activity prediction problem and has also been applied to other problems such as image classification and retrieval [2], stock market forecasting [3] and text classification [4].

Current MIL algorithms, that basically extend SL concepts for dealing with MIL problems, are Diverse Density (DD) [3], Expectation-Maximization Diverse Density (DD-EM) [5], Axis Parallel Concepts (APR) [1], Citation KNN (K-Nearest Neighbors) [6], and variations of Support Vector Machine (mi-SVM and MI-SVM) [4]. In this paper a new MIL filter method that is based on the Citation KNN is proposed. The principle of the proposed method is based on the homogeneity of negative bags and on the assumption of coherence of their spatial distribution. Assuming such a spatial coherence, the negative samples within positive bags are expected to be spatially related to the negative samples within negative bags, so the former can be filtered from positive bags by means

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

of a distance classifier. At the end of filtering both bags are expected to be homogeneous. Filtering reduces the spatial shift of positive bags and paves the way to a more consistent bag classification with methods such as Citation KNN.

The remainder of this paper is organized as follows. Section 2 summarizes previous works on MIL. Next, section 3 describes the proposed homogeneity bag operator. Section 4 presents numerical experiments comparing the proposed approach with existing MIL classifiers. Finally, discussions and conclusions are presented in Section 5.

2 Related Work

In the next subsections, current approaches to MIL are described.

2.1 Axis Parallel Concepts (APR)

Axis Parallel Hyper Rectangle was the first class of algorithms proposed to solve MIL problems. This algorithm was originally proposed by Dietterich *et al.* [1] and applied to the drug activity problem. Its main idea is to find an *axis-parallel hyper-rectangle* (APR) in feature space that bounds the positive samples. The APR must contain at least one instance of each positive bag, meanwhile all instances of negative bag are outside.

2.2 Diverse Density (DD)

Diverse Density (DD) was originally described by Maron & Ratan [3]. The main idea of DD is to find a concept point in feature space that is near at least one instance from each positive *bag* and simultaneously far from the instances from negative *bags*. In [7], the optimum concept point is defined according to how many positive *bags* have instances near the concept point, and how far of the negative instances from that point. According to [8], the optimum concept point can be found by maximizing the following DD function:

$$arg\ max_x \prod_i Pr(x = t|B_i^+) \prod_i \Pr(x = t|B_i^-) \tag{1}$$

where B_i is the i^{th} bag of the data set B, and a *positive bag* is represented by B_i^+ and a *negative bag* B_i^-.

2.3 Expectation-Maximization Diverse Density (EM-DD)

EM-DD originally proposed in [5] is an extension of the standard DD algorithm. Although EM-DD is based on the same theoretical baseline of DD, it adopts a different approach to find the most likely concept point. EM-DD combines Expectation-Maximization (EM) with the original DD. Seed concept point is selected according to the original DD, followed by Expectation and Maximization iterations aiming at maximizing EM-DD objective function. The algorithm iterates until convergence.

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

2.4 Support Vector Machines (mi-SVM and MI-SVM)

Many MIL algorithms are adaptations of conventional SL algorithms that are extended for bag classification. In the cases of MI-SVM and mi-SVM [4], margin maximization is redefined in order to consider MIL constraints. In fact, MI-SVM deals with the problem at bag level, whereas mi-SVM deals with sample level. The bag margin is defined according to the positive samples within the positive bag. Since our method is aimed at bag level, our results are compared with MI-SVM.

2.5 Citation kNN

Citation kNN is based on the nearest neighbours rule and in the concepts of citation and reference [6]. A bag is labelled according not only to its neighbours, but also according to the bags that have the reference bag as a neighbour. The key intuition behind Citation kNN for MIL problem is how to transform the measure between instances (such as in standard kNN) in a measure between bags. In [6], it is proposed the use of the Housdorff Distance to measure a distance between two subsets of instance. Following the definition in [6] two sets of points A and B are within Housdorff distance d of each other if and only if every point of A is within distance d of at least one point of B, and vice versa.

The authors claims that only with the use of the Hausdorff distance is not enough to obtain promising results, and other improvements are required. So the authors proposes the use of the notion of citation from library. In this case, besides consider the bags as the nearest neighbours (references) of some bag B_i, it is also considered the bags that recognize B_i as their neighbours (citers).

3 Homogeneity Bag Operator

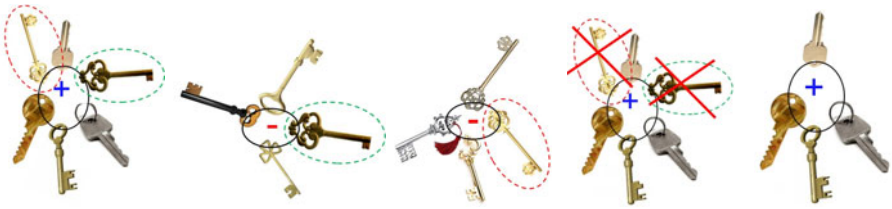
In favor of introducing the proposed homogeneity bag operator, a classical MIL problem is considered as an illustrative example. This problem was adapted from [9].

Imagine that in a company there are several staff members, and each one has a key chain with many different keys. Only some of the staff members have the key to enter a certain room. The task is to predict whether a certain key or certain key chain grants access to this room. To solve this problem, one needs to find the key that all "positive" key chains have in common.

This searching process can be optimized by applying a pre-processing step comparing keys from "negative" key chains with the ones from "positive" key chains. It is known that within a "negative" key chain there are only "negative" keys. Thus, the process can be simplified by firstly removing keys from "positive" key chains similar with keys from "negative" ones.

Figure 1 illustrates the proposed homogeneity bag operator for the illustrative problem. Figure 1(a), 1(b) and 1(c) depict three different key chains (bags). The first one contains the desired key (positive bag) and the remaining ones have only keys that do not open the desired room, i.e., two negative bags. Figure 1(d)

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.



(k.c. 1) (a) (k.c. 2) (b) (k.c. 3) (c) (k.c. 1 original) (d) (k.c. 1 final) (e)

Fig. 1: Illustration of the proposed homogeneity bag operator: (a) positive bag; (b) and (c) negative bags; (d) negative keys to be eliminated from key chain 1 ; and (e) resulting key chain 1

illustrates keys within the positive bag similar with ones from negative bags. Finally, Figure1(d) exhibits the filtered positive bag.

The proposed homogeneity bag operator can be used to improve the Citation K-NN algorithm. This classifier decides whether a bag is positive or negative using a bag distance measure based on similarities between instances of two bags. By removing negative instances from positives bags, this bag distance measure can be improved.

3.1 Similarity Measure and The Homogeneity Operator

In order to filter negative instances from positive bags, a instance similarity measure must be defined. The similarity between an instance j in the i^{th} positive bag B_{ij}^{+} and an instance l in the k^{it} negative bag B_{kl}^{-} is defined as the number of attributes "close" enough to each other, *i.e.*,

$$Sim(B_{ij}^{+}, B_{kl}^{-}) = \sum_{m=1}^{n_k} Close(B_{ijm}^{+}, B_{klm}^{-}) \tag{2}$$

in which B_{ijm}^{+} and B_{klm}^{-} correspond to the m^{th} attribute of the two instances considered, $Close(B_{ijm}^{+}, B_{klm}^{-}) = 1$, if $|B_{ijm}^{+} - B_{klm}^{-}| < \alpha |B_{ijm}^{+}|$, and 0 otherwise. n_k is the number of attributes and α is a parameter controlling the proximity degree of instances.

Once the similarity is computed, in order to decide if the instance j of the positive bag should be removed a threshold β is applied to the similarity value (2). If the similarity value is greater than β the instance is classified as an negative instance within a positive bag and is removed in order to promote bag homogeneity.

This procedure is repeated for each instance in all positive bags. If all instances of a given bag are selected to be removed, the one with the lowest similarity is kept to avoid generating an empty positive bag.

The values of α and β are essential to define the similarity measure and the homogeneity operator, respectively. In this work, these parameters were

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

empirically set to $\alpha = 0.15$ and $\beta = 0.85$ for all experiments.

4 Numerical Experiments

In this section the proposed approach is evaluated using popular benchmark MIL datasets. Four datasets were considered in the experiments namely *MUSK1*, *Elephant*, *Tiger* and *Fox*. In *MUSK1* dataset [1] the task is to predict drug activity from structural information of a molecule. Each molecule is represented by a bag and each instance within a bag represents a distinct structural configuration of the molecule. In *Elephant*, *Tiger* and *Fox* datasets [4] the objective is to differentiate images containing elephants, tigers and foxes from those that do not. Each bag represents an image and each instance of a bag is a region of interest within the image. Table 1 details each dataset considered.

Table 1: MIL Dataset

	Bags			Number of	Number of
Dataset	Positive	Negative	Total	Features	Instances
MUSK 1	47	45	92	166	476
Elephant	100	100	200	230	1391
Fox	100	100	200	230	1220
Tiger	100	100	200	230	1320

The proposed homogeneity bag operator was used as a pre-processing step for the Citation K-NN algorithm, removing possible negative instances from training set positive bags. The results achieved by this approach were then compared with classic MIL classifiers, including Citation K-NN without the pre-processing step.

All experiments were performed using the Multiple Instance Learning Library (MILL) [7]. MIIL is an open-source toolkit for multiple instance learning, developed in MatLab, and it is largely used.

Table 2 presents results for the proposed approach and all other MIL methods discussed in section 2. Leave-one-out cross-validation was used to compute the accuracy of all classifiers.

Table 2: Results

Dataset	MUSK1	Elephant	Tiger	Fox
DD	86,6%	78,5	75,0%	62,0%
EM-DD	86,9%	76,0%	74,5%	57,0%
Citation k-NN	90,2%	77,2%	76,5%	54,0%
Citation k-NN with Elimination	**91,3**%	**80,1**%	**81.0**%	**56,5**%
MI-SVM	75,0%	78,0%	82,5%	47,0%
APR	86,0%	71,0%	55,0%	58,0%

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

5 Discussions and Conclusions

The proposed homogeneity bag operator increased the accuracy of Citation K-NN for all datasets. Furthermore, the proposed method achieved the best results in 2 out of 4 datasets when compared with classic MIL classifiers.

The improvement observed for Citation-KNN was expected, since by correctly removing negative instances from positive bags, the accuracy of the Housdorff Distance is improved. This distance measure is used by the classifier to estimate similarities between bags and, consequently, to decide whether a bag is positive or negative.

Future work shall address use of alternative distance metrics for similarity and development of heuristics to automatically tune the parameters of the proposed homogeneity bag operator.

6 Acknowledgments

This work has been supported by the Brazilian agency CAPES.

References

[1] T. G. Dietterich, R. H. Lathorp, and T. Lozano-Perez. Solving the multiple instance problem with axis-parallel retangles. *Artificial Intelligence*, 89(1-2):31–71, 1997.

[2] O. Maron and A.L. Ratan. Multiple-instance learning for natural scene classification. In *International Conference on Machine Learning (ICML)*, pages 341–349, 1998.

[3] O. Maron and T. L. Perez. A framework for multiple-instance learning. In *Advances in Neural Inf. Proc. Systems (NIPS)*, pages 570–576, 1998.

[4] S. Andrews, I. Tsochantaridis, and T. Hofmann. Support vector machines for multiple-instance learning. In *Advances in Neural Inf. Proc. Systems (NIPS)*, pages 561–568, 2003.

[5] Qi Zhang and Sally A. Goldman. EM-DD: An improved multiple-instance learning technique. In *Advances in Neural Inf. Proc. Systems (NIPS)*, pages 1073–1080, 2001.

[6] J. Wang and J. D. Zucker. Solving the multiple instance problem: A lazy learning approach. In *International Conference on Machine Learning (ICML)*, pages 1119–1126, 2000.

[7] J. Yang. MILL: A multiple instance learning library. `http://www.cs.cmu.edu/~juny/MILL`, 2008. last visit on September 2013.

[8] Y. Chen, J. Bi, and J. Z. Wang. MILES: Multiple-instance learning via embedded instance selection. *IEEE Trans. PAMI*, 28(12):1931–1947, 2006.

[9] B. Babenko. Multiple instance learning: Algorithms and applications. Technical report, University of California San Diego, San Diego, 2008. available at http://vision.ucsd.edu/ bbabenko/data/bbabenko_re.pdf.