CS395T Data Mining
Project report
One-class SVM formulations for Multiple Instance learning

Sudheendra Vijayanarasimhan

1 Introduction

Multiple Instance learning (MIL) considers a particular form of weak supervision in which the learner is given a set of *positive* bags and *negative* bags. *Positive* bags are sets of instances containing atleast one positive example and *negative* bags are sets of instances all of which are negative. A number of binary SVM based solutions have been proposed to this problem like the Normalized Set Kernel of Gartner et. al, 2002 ([1]) which represents the bag as the sum of all its instances normalized by its 1 or 2-norm and the sparse MIL (sMIL) technique of Razvan and Mooney, 2007 ([2]) which improves upon NSK by considering a weaker balancing constraint. In this project I plan to look at equivalent formulations for a one-class SVM and empirically evaluate if ignoring the negative bags in the formulation is detrimental to the solution found.

2 Related Work

A number of 2-class SVM based formulations have been looked at in the literature. The following are a few relevant MIL SVM formulations

- Normalized Set Kernel (NSK)

  In the Normalized Set Kernel of Gartner et. al, 2002 ([1]) a bag is represented as the sum of all its instances, normalized by its 1 or 2-norm. The resulting representation is then trained using a standard SVM. The formulation for NSK is as follows

- sparse MIL (sMIL)

  The sparse MIL formulation of Razvan and Mooney, 2007 ([2]) considers the equation for the positive bags in the formulation of NSK as a balancing constraint. The balancing constraint of NSK is too strong since it assumes that all the instances in a positive bag are positive. Since this is problematic when the positive bag is particularly sparse in positive examples they consider the constraint that expresses that *at least* one instance from the bag is positive. The formulation for the same is as follows

- MI-SVM

$$
\begin{aligned}
\text{minimize:} \quad & \tfrac{1}{2}||w||^2 + \tfrac{C}{|\mathcal{X}_n|} \textstyle\sum_{x \in \mathcal{X}_n} \xi_x + \tfrac{C}{|\mathcal{X}_p|} \textstyle\sum_{X \in \mathcal{X}_p} \xi_X && (1) \\
\text{subject to:} \quad & max_{x \in X} \;\; y(w\phi(x) + b) \geq 1 - \xi_X, \forall X \in \mathcal{X}_p, X_n && (2)
\end{aligned}
$$

  This is the maximum bag margin formulation of Andrews et. al (2003) [3]. The associated heuristic algorithm starts by training a standard SVM. This is followed by relabeling of instances in positive bags using the decision hyperplane. If a positive bag contains no instances that are positive according to this hyperplane then the instance with the maximum value of the decision function is relabeled as positive and the SVM is retrained on this relabeled data. This is continued till there are no more labels to be changed.

- A regularization framework for Multiple-Instance learning In the method proposed by Cheung and Kwok (2006) [4] a loss function is introduced between the label of a bag and the label of the most positive instance in the bag and SVM is formulated by including this loss function in the objective. This is based on relaxing the idea of mi-SVM that the label of a positive bag is equal to the label of its most positive instance. But the objective function is no longer convex because of the max function used for the loss and therefore they directly formulate the dual problem instead of the primal and solve it using CCCP([5]).

Positive bags are easily constructed in all of the above cases. For example, in image retrieval each returned set of images can be considered as a positive bag, segmentation where each image is a positive bag containing atleast one valid segmentation. On the other hand it is not clear on how to choose negative bags in these cases and they are typically constructed from examples that are known to be non-positive.

Ray and Craven (2005) ([6]) observe that the nature of the negative instances in the positive bags may be different from the nature of the negative instances in the negative bags. If this is the case then one would be dealing with 3 different distributions which might or might not be separable using the single hyperplane found by all of these methods. And so it appears that most of these methods work well only when the negatives in the positive bags are similarly distributed to the negatives in the negative bags ([2]).

Therefore even though we've the freedom of constructing the negative bags from any set of instances that is not positive the most gains are obtained when these are sampled from a distribution similar to that of the negatives in the positive bags. This might not be possible in some cases like image retrieval where no information is known about the distribution of the noisy images in each retrieved set.

In such situations completely ignoring the negative bags in the formulation and considering only the positive bags and using clustering techniques or a one-class SVM might be fruitful. In the following work we will formulate a one-class SVM for the MI problem and compare it with the standard one-class SVM on an image dataset. We will also compare the one-class SVM solutions with the 2-class methods outlined above to study the effect of ignoring the negative bags.

3 A one-class SVM approach to MIL

A one-class SVM is a function f that takes the value +1 in a "small" region capturing most of the data points and -1 elsewhere. One-class SVMs are typically used for novelty detection where the task is to say whether a new example is unlike any one of training examples. One-class SVMs have also been applied to the task of unsupervised learning for character regognition ([7]).

The MI problem could be solved using the one-class SVM by simply considering all instances in positive bags as unlabeled data and then estimating a function that returns +1 in a "small" region that should correspond to the true positives. The function is found by mapping the data into feature space corresponding to a kernel and then separating them from the origin with maximum margin. This corresponds to the following quadratic problem

$$\begin{aligned} \text{minimize:} \quad & \tfrac{1}{2}||w||^2 + \tfrac{1}{\nu l}\textstyle\sum_i \xi_i - \rho && (3) \\ \text{subject to:} \quad & (w.\phi(x)) \geq \rho - \xi_i, \xi_i \geq 0. && \\ & && (4) \end{aligned}$$

Here l is the total number of training examples. But the above formulation does not respect the MI constraint which states that positive bags should contain atleast one positive instance. From Ray and Craven (2005) [6] it is clear that even though ignoring the bag constraint and solving the standard supervised problem produces results comparable to MI methods in most datasets, when the bags are very sparse MI methods invariably perform better.

As seen in Section 2 the MI constraint can be captured in a number of ways. Of these the idea of Andrews et. al [3] is closest to capturing the MI constraint since it states that the maximum value of y within a bag should be greater than +1 if the bag is positive. But even though the max function is convex it is not smooth and so the standard quadratic optimization techniques cannot be applied. Therefore we will apply the technique used in [4]. The one-class MI-svm formulation is given in Figure 1

Here l denotes the total number of instances within all bags while n denotes the total number of bags. The penalty for bags and instances have been seperated out because the bag constraint is a stronger constraint as we want atleast

$$\begin{array}{rc} \text{minimize:} & \frac{1}{2}||w||^2 + \frac{1}{\nu l}\sum_i \xi_i + \frac{1}{\nu n}\sum_i \Xi_i - \rho \quad (5) \\ \text{subject to:} & (w.\phi(x)) \geq \rho - \xi_i, \xi_i \geq 0. \\ & max_{x \in X} \quad (w\phi(x)) \geq \rho - \Xi_X, \forall X \in \mathcal{X}_p, \Xi_X \geq 0 \end{array} \quad (6)$$

Figure 1: one-class MI-svm formulation

one instance to be positive in each bag. On the other hand individual instances might not be all positive and therefore the penalty should be lower.

Using multipliers $\alpha_i, \beta_i \geq 0$, we introduce a Lagrangian

$$\begin{array}{rl} L(w,\xi,\Xi,\rho,\alpha,\beta) = & \frac{1}{2}||w||^2 + \frac{1}{\nu l}\sum_i \xi_i + \frac{1}{\nu n}\sum_i \Xi_i - \rho - \sum_i \alpha_i((w.\phi(x_i)) - \rho + \xi_i) \\ & -\sum_i \alpha_i(max_{x \in X_i}(w.\phi(x)) - \rho + \Xi_i) - \sum_i \beta_i \xi_i - \sum_i \beta_i \Xi_i \\ \alpha \geq & 0 \\ \beta \geq & 0 \end{array} \quad (7)$$

Now, because of the presence of the *max* function the lagragian is not differential. But by using the sub-gradient of the *max* function and setting the derivatives to zero we can obtain the dual of the above problem. For the pointwise maximum function $h(x) = max_{1 \leq i \leq p} h_i(x)$ its subdifferential at x_0 is the convex hull of the union of subgradients of "active" functions at x_0. Function h_i is said to be active if $h_i = max_{1 \leq i \leq p} h_i(x)$. Introducing variables a_{ij} to denote whether $(w.\phi(x_j))$ is active or not in the max function yields the solution

$$\begin{array}{rll} w = & \sum_i \alpha_i \phi(x_i) + \sum_i \alpha_i (\sum_j a_{ij}\phi(x_j)) & \\ \alpha_i \leq & \frac{1}{\nu l}, & when\ i\ is\ an\ instance \\ \alpha_i \leq & \frac{1}{\nu n}, & when\ i\ is\ a\ bag \\ \sum_i \alpha_i = & 1. & \\ \sum_j a_{ij} = & 1, & \forall i \\ a_{ij} = & \frac{1}{n_a}, & if\ (w.\phi(x_{ij})) = (max_{x \in X_i}(w.\phi(x)), x_{ij} \in X_i \\ = & 0, & otherwise \end{array} \quad (8)$$

We initialize $a_{ij}^{(0)} = 0$ for all bags, and the a_{ij}s are updated as $a_{ij} = 0, if x_j$ is not active in the max function and $a_{ij} = 1/n_a$ if it is. Here n_a denotes the

```
initialize $a_{ij}^{(0)} = 0, \forall x_{ij} \in X_i \in \chi_p$
while any one of the max constraints is violated do
    compute the one-class SVM solution $w, \rho$ using $a_{ij}$;
    foreach positive bag $X_i \in \chi_p$ do
        compute outputs $f_x =< w, x > +\rho, \forall x \in X_i$ ;
        let $f_{max}$ be the largest value of $f_x$ ;
        and $n_a$ denote the number of $x$s with $f_x = f_{max}$ ;
        foreach instance $x_{ij} \in X_i$ do
            if $f_x == f_{max}$ then
                $a_{ij} = \frac{1}{n_a}$ ;
            else
                $a_{ij} = 0$ ;
            end
        end
    end
end
OUTPUT $(w, \rho)$
```

Figure 2: The psuedo-code for the one-class MI-SVM

number of active instances. The pseudo-code for this procedure is as shown in Figure 2.

4 Datasets

To evaluate the proposed one-class MIL method two different datasets from the computer vision were considered. The following is a brief description of the same.

- **SIVAL dataset**

  The SIVAL dataset contains segmented images of various objects in different scenes. A positive instance is a segment containing the object, while all others are negative. An image (bag) is labeled positive if it contains the object. The classification task is to say whether a given *segment* is positive or not.

  The SIVAL images contain different objects in very similar scenes. Therefore, in this dataset there is a reasonable correlation between the negative segments found in *positive bags* and the negative segments found in the *negative bags.*

- **Google dataset**

  The Google dataset contains images obtained from keyword searches of a particular category's name. Airplanes, Cars, Faces, Guitars, Leopards,

Category	AUROC - training		AUROC - test	
	one-class	one-class MI-svm	one-class	one-class MI-svm
ajaxorange	64.22	**65.40**	63.72	**64.14**
apple	50.64	**50.96**	49.70	49.62
banana	63.16	**64.82**	61.24	**63.20**
bluescrunge	49.00	**52.44**	47.86	**50.50**
candlewithholder	76.14	**76.62**	77.22	77.20
cardboardbox	77.62	**78.84**	75.80	**77.00**
checkeredscarf	78.32	78.06	78.62	78.42
cokecan	76.78	74.32	76.00	72.98
dataminingbook	81.36	80.84	77.16	76.34
dirtyrunningshoe	76.54	**77.70**	71.86	**72.60**
dirtyworkgloves	74.82	**75.90**	77.78	**78.74**
fabricsoftenerbox	83.62	**83.76**	82.42	82.24
feltflowerrug	60.56	57.00	56.66	52.18
glazedwoodpot	48.52	**48.84**	44.08	44.06
goldmedal	49.66	**55.78**	49.86	**55.24**
greenteabox	64.36	**64.92**	63.76	**64.28**
juliespot	50.86	**52.00**	47.68	**48.58**
largespoon	69.94	**72.86**	75.74	**77.72**
rapbook	69.62	**71.32**	72.22	**73.50**
smileyfacedoll	56.50	56.20	59.80	**59.98**
spritecan	68.92	**69.04**	67.82	67.22
stripednotebook	86.06	**87.84**	82.72	84.36
translucentbowl	43.06	**44.18**	44.48	**46.24**
wd40can	69.78	66.02	65.68	60.64
woodrollingpin	78.62	**79.20**	76.94	76.50
Average	66.75	**67.39**	65.87	**66.14**

Table 1: Area under the ROC curve for different categories in the SIVAL dataset on both training and test data averaged over 5 random trials. Numbers highlighted in bold area cases where adding the MI constraint improves the area under the ROC. We can clearly see that there is an improvement for majority of the categories even though overall average is only slightly larger.

Category	AUROC - training		AUROC - test	
	one-class	one-class MI-svm	one-class	one-class MI-svm
airplane	71.10	**71.96**	92.46	**95.60**
cars_rear	69.08	67.72	89.52	88.18
face	63.18	63.18	76.26	76.26
guitar	40.42	39.76	69.64	70.00
leopard	69.20	68.40	91.62	91.48
motorbike	68.88	68.98	86.90	**87.18**
wrist_watch	63.42	63.14	79.26	79.54
Average	63.61	63.31	83.67	**84.03**

Table 2: Area under the ROC curve for different categories in the Caltech-7 dataset on both training and test data averaged over 5 random trials. Numbers highlighted in bold are cases where the one-class SVM is better than the two-class version.

Motorbikes and Wristwatches are the 7 different categories available in the Google dataset. A large number of images for a category have been downloaded and randomly assigned into bags of size 25. Negative bags are constructed from Caltech Background images.

The test set was constructed from the Caltech dataset of the category against Background images. The classification task is say whether a particular image belongs to the category of the background.

Since the images obtained from keyword search returns can contain images of synonyms of the category and other irrelevant images which might not be available in the Caltech Background category the distribution of negative images in *positive bags* and those in *negative bags* might be completely different.

5 Experiments and Results

Two different sets of experiments were conducted using the above datasets. First the proposed one-class MI method was evaluated on both the datasets and the one-class MI method was compared with the two-class NSK-SVM approach to MIL to evalute if ignoring the negative bags in the formulation simplifies the problem and if so under what conditions.

5.1 One-class MI-SVM

Since we do not use the labels on positive instances in either the standard one-class SVM or the one-class MI-svm both the results on the training data and the test data are equally relevant here. Both the SIVAL and Google datasets were randomly split into 5 runs each containing a training set of 20 images ($\tilde{2}$0*30

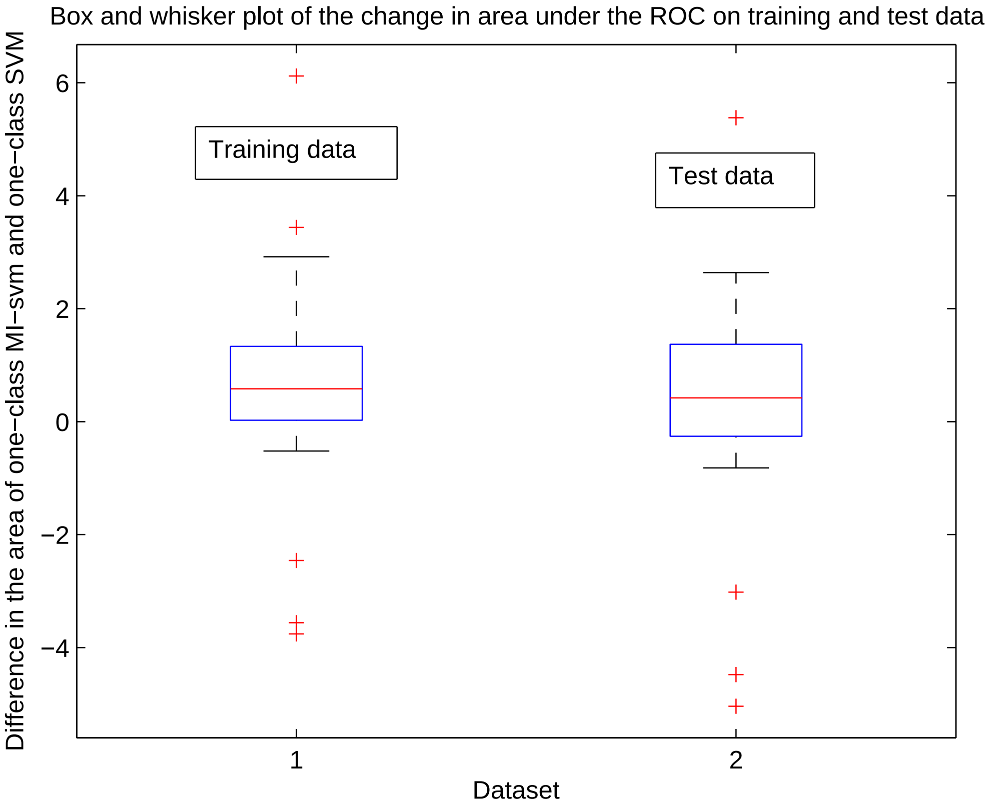

Figure 3: A box and whisker plot of the difference in the area under ROC for the two methods. The mean and lower quartile are above zero for both test and training data implying that the MI constraint does improve the area under the ROC for most categories

segments for SIVAL, 20*25 images for Google). All results are averaged over the 5 runs.

Kernel and other parameters were optimized separately and the same value was used for both the standard SVM and the MI-svm since only the comparison on the same set of parameters would be relevant. The results for the SIVAL dataset are for a quadratic kernel with a coefficient of $5 * 10^{-6}$ and $\nu = 0.9$. For the Google dataset the RBF kernel/quadratic kernel was used based on accuracy on a held-out set with $\gamma = 5 * 10^{-6}$ and $\nu = 0.9$.

Table 1 shows the Area under the ROC curve for the classification task of predicting whether a *segment* is positive or not. The first two columns are on the training data while the last two are on the test data. Numbers highlighted in bold refer to cases where the MI method did better than the standard one-class

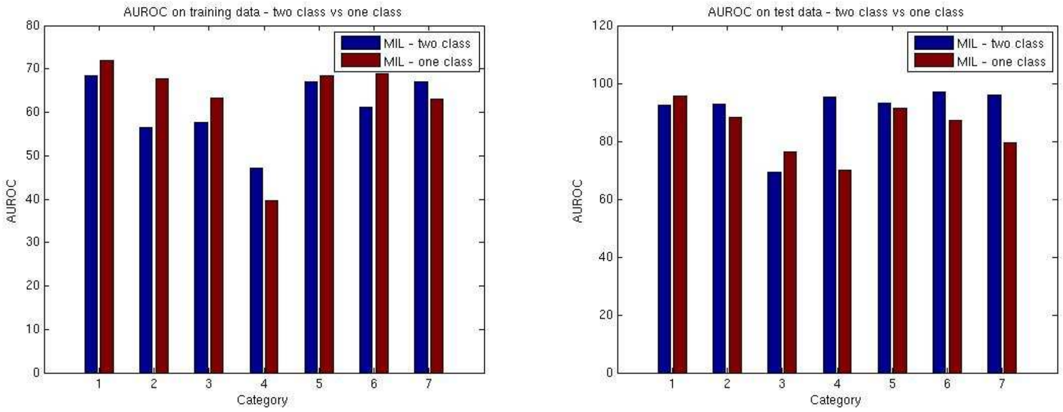

Figure 4: A box and whisker plot of the difference in the area under ROC for the two methods. The mean and lower quartile are above zero for both test and training data implying that the MI constraint does improve the area under the ROC for most categories

method. The MI method performs better than the standard one-class method in a majority of the categories with a maximum improvement of 3.44 in the case of *bluescrunge* on the training set. But the overall averges differ by less than 1 point and it cannot be considered as definitive that the MI method is better than the one-class. Figure 3 shows a whisker and box plot of the difference in the area under the ROC for the two methods. We see that there are a number of negative outliers which could be responsible for the low average improvement.

Table 2 shows the same numbers on the Google dataset. There is a slight improvement on the test images for the MI method as opposed to the standard one-class method.

5.2 One-class MI-SVM vs NSK-SVM

To empirically evaluate the effect of leaving out *negative bags* in the formulation the proposed one-class method was compared with the Normalized Set Kernel (Section 2) for two-class classification. The classification tasks for both the datasets are as explained in the previous experiment. For the SIVAL dataset the NSK was used in conjunction with a quadratic kernel with a coefficient of $5 * 10^{-6}$ and $\nu = 0.2$ and for the Google dataset the RBF kernel was used with $gamma = 5 * 10^{-6}$ and $\nu = 0.5$.

Table 3 shows the area under the ROC curve for the two methods on both training and test data on the SIVAL dataset. The one-class method is better than the NSK in 10 out of 25 categories. But the overall average for the one-class method is much lower than the average for the NSK. As noted in Section 4, in the SIVAL dataset negative segments in the positive bags are similar in distribution to the negative segments in the negative bags. So the additional

Category	AUROC - training		AUROC - test	
	NSK	one-class MI-svm	NSK	one-class MI-svm
ajaxorange	77.74	65.40	73.56	64.14
apple	72.46	50.96	67.44	49.62
banana	76.98	64.82	69.78	63.20
bluescrunge	82.46	52.44	74.20	50.50
candlewithholder	68.50	**76.62**	63.76	**77.20**
cardboardbox	67.22	**78.84**	63.98	**77.00**
checkeredscarf	79.20	78.06	75.50	**78.42**
cokecan	80.58	74.32	77.50	72.98
dataminingbook	74.74	**80.84**	64.86	**76.34**
dirtyrunningshoe	76.68	**77.70**	73.14	72.60
dirtyworkgloves	65.32	**75.90**	63.36	**78.74**
fabricsoftenerbox	89.10	83.76	86.42	82.24
feltflowerrug	78.12	57.00	75.96	52.18
glazedwoodpot	68.66	48.84	65.52	44.06
goldmedal	74.78	55.78	63.44	55.24
greenteabox	83.16	64.92	78.34	64.28
juliespot	74.26	52.00	66.10	48.58
largespoon	69.14	**72.86**	63.76	**77.72**
rapbook	65.38	**71.32**	57.72	**73.50**
smileyfacedoll	78.72	56.20	72.08	59.98
spritecan	76.90	69.04	68.72	67.22
stripednotebook	83.44	**87.84**	77.30	**84.36**
translucentbowl	71.30	44.18	73.62	46.24
wd40can	81.56	66.02	73.70	60.64
woodrollingpin	63.68	**79.20**	58.98	**76.50**
Average	75.20	67.39	69.95	66.14

Table 3: Area under the ROC curve for different categories in the SIVAL dataset on both training and test data averaged over 5 random trials. Numbers highlighted in bold are cases where the one-class SVM is better than the two-class version.

data available with the NSK in the form of negative bags seems to help in learning to differentiate between positive and negative segments.

On the other hand Figure 4 shows a bar graph of the area under ROC for the two methods on the Google dataset. The figure on the left is the result on the training images (the noisy google images containing both the category and other images) while on the right we've the result on the Caltech dataset.

The one-class method is much better than the NSK at identifying correctly the positive class from the noisy training images. Considering the fact that in the Google dataset the negatives in the positive bags and negative bags might be from different distributions this seems to suggest that the one-class method might be advantageous when the two distributions are dissimilar. This is because in this situation the NSK would have to distinguish between 3 different classes using a single hyperplane.

But on the test images the NSK is much better. This is because the negative test images consisting of Caltech background images are available to the NSK during training as negative bags as opposed to the one-class method which trains only on positive bags. In this situation extracting the positive class from the positive bags and training a standard two-class SVM with background images should improve results on the test data.

6 Conclusion

A one-class SVM was formulated for the MI problem and it was compared with a standard one-class SVM for two different image datasets. Results on the two datasets suggest that including the MI constraint does improve classification for most cases but not by much. The iterative procedure outlined in Section 3 reduces the objective in subsequent iterations but for most of the classes the constraint gets satisfied in under 2-3 iterations. Since the procedure is based on a sub-gradient method it finds the optimal solution for the given formulation.

But based on the results using ground truth labels it was found that the solution found is not optimal for the actual MI problem. It was also observed that the active instances in some bags corresponded to the negative class which could have resulted in the negative change in the AUROC for some categories. These results seem to suggest that even though the max constraint improves on the standard one-class SVM it does not completely capture the Multiple Instance Learning criteria. Transductive constraints on maximally separating instances within positive bags would be something to look at in the future.

The one-class SVM was also compared with a standard two-class MIL method, the NSK. Based on the results on two different datasets it was found that it might be advantageous to use the one-class method when the distribution of the negative instances in *positive* and *negative* bags are different. This is because in such situations the NSK might be trying to seperate three different distributions (positives, negatives in positive bags and negatives in negative bags) using a single hyperplane which might not be feasible. Extracting the positives from the positive bags using the one-class method and then constructing a standard two-class SVM might lead to better classification in such situations.

References

[1] Gartner, T., Flach, P., Kowalczyk, A. & Smola, A. (2002). Multiple Instance kernels. *In Proceedings of the 19th International Conference on Machine Learning*

[2] Razvan Bunescu, Ray Mooney (2007). Multiple Instance Learning for Sparse Positive Bags.

[3] Andrews, S., Tsochantaridis, I., & Hofmann, T. (2003). Support Vector Machines for Multiple Instance Learning. *Advances in Neural Information Processing Systems 15. Cambridge, MA: MIT Press.*

[4] Pak-Ming Cheung and James T. Kwok (2006) A Regularization Framework for Multiple Instance Learning. *ICML '06: Proceedings of the 23rd international conference on Machine learning*

[5] Yuille, A., & Rangarajan, A. (2003). The Concave Convex Procedure. *Neural Computation, 15, 915*

[6] Ray, S., and Craven, M. (2005). Supervised versus Multiple Instance Learning: An Empirical Comparison. *Proceedings of 22nd International Conference on Machine Learning (ICML-2005) (pp. 697Bonn, Germany.*

[7] B. Scholkopf and J. Platt and J. Shawe-Taylor and A. Smola and R. Williamson. Estimating the Support of a High-dimensional Distribution. Technical Report 99-87, Microsoft Research, 1999.