Accepted in 23rd International Conference on Pattern Recognition (ICPR), 2016

Witness Identification in Multiple Instance Learning Using Random Subspaces

Marc-André Carbonneau, Eric Granger and Ghyslain Gagnon
École de technologie supérieure, Université du Québec, Montréal, Canada
marcandre.carbonneau@gmail.com, eric.granger@etsmtl.ca, ghyslain.gagnon@etsmtl.ca

***Abstract*—Multiple instance learning (MIL) is a form of weakly-supervised learning where instances are organized in bags. A label is provided for bags, but not for instances. MIL literature typically focuses on the classification of bags seen as one object, or as a combination of their instances. In both cases, performance is generally measured using labels assigned to entire bags. In this paper, the MIL problem is formulated as a knowledge discovery task for which algorithms seek to discover the witnesses (i.e. identifying positive instances), using the weak supervision provided by bag labels. Some MIL methods are suitable for instance classification, but perform poorly in application where the witness rate is low, or when the positive class distribution is multimodal. A new method that clusters data projected in random subspaces is proposed to perform witness identification in these adverse settings. The proposed method is assessed on MIL data sets from three application domains, and compared to 7 reference MIL algorithms for the witness identification task. The proposed algorithm constantly ranks among the best methods in all experiments, while all other methods perform unevenly across data sets.**

***Index Terms*—Multiple Instance Learning; Random Subspace Methods; Witness Identification; Knowledge Discovery.**

I. Introduction

In multiple instance learning problems, instances are grouped in bags, and a label is provided for the whole bags. The individual labels of the instances are unknown. The standard formulation of MIL assume negative bags do not contain positive instances, while positive bags are said to contain at least one positive instance, called witness [1].

MIL have been successfully applied to various applications, such as molecule conformation classification [2] and content-based image retrieval (CBIR) [3]–[5]. More recently, MIL algorithms attracted attention in the medical community, especially for computer-aided diagnostic from images [6]–[8] because it allows learning from loosely annotated images.

In some applications, phenomenons are quantified using a set of observations. Identifying the truly informative instances, the witnesses, helps researchers better understand the phenomenon. For example, Palachanis [9] uses MIL to identify the genomic features governing the bonding of transcription factors in gene expression. In this case, bags represent genes, and transcription factors are instances. Witnesses are identified, and found to be corresponding to biological observations. In automated personality assessment from speech signals, data sets are created by psychologists that assign personality traits labels to whole speech segments. These experts perform this task intuitively, and thus, it is not clear what parts of the signal provided relevant cues for classification [10]. Being able to identify witnesses from positive bags could provide insight on the nature of data. As another example, by comparing the social media posts that a user either reads or ignores, one could infer user-specific elements of interest. All these cases correspond to the identification of witnesses in MIL data sets, which is more of a knowledge discovery task than a classification task.

Not all MIL algorithms allow to classify instances instead of bags. Many MIL algorithms based on bag distance measures [11], [12] and bag embedding [13], [14] do not provide information at instance-level, and therefore cannot directly be used in witness identification problems. However, some of these methods, like MILES [5] and Citation-kNN [15], can be adapted for the task. In contrast, instance-based MIL methods like axis parallel rectangle (APR) [2], mi-SVM, MI-SVM [3] and KI-SVM [4] infer bag labels based on individual instance classification, and thus can be used directly for witness identification. Although these methods can achieve a high level of performance in specific situations, they often perform poorly when the proportion of positive instances in positive bags, hereafter called the witness rate (WR), is low. In other cases, the methods cannot deal with witnesses sampled from multimodal positive data distributions. The modes of the distributions are clusters corresponding to latent variables in the data set, which will hereafter be called concepts.

In this paper a new method named Random Subspace Witness Identification (RSWI) is proposed. A related method was used in [16] to design MIL ensembles for classification, and was shown to be robust to both low WR and multi-concept problems. RSWI computes a score for each instance that corresponds to its likelihood of being a witness. To compute these scores, all instances of the data set are projected in several random subspaces. Clustering is performed in each subspace, and the proportion of instances belonging to positive bags in each cluster is computed. The score of an instance is obtained by adding these proportions for each cluster it was assigned to. The random subspaces help capture relations in the data and provide robustness against the effects of irrelevant and redundant features, especially when using distance-based clustering methods like k-means with Euclidean distance.

To validate RSWI, the performance of several MIL algorithms with witness identification capabilities are compared and analyzed. Since witness identification is an aspect that has not yet been deeply explored, most existing MIL data sets do not include instance-level annotation. Thus, 2 new data sets have been created using data from real-world applications. The data sets were made publicly available by the authors on his personal website (https://sites.google.com/site/marcandrecarbonneau/).

II. Witness Identification in MIL Methods

Several instance-based MIL methods have been proposed for MIL. Instance-based methods classify instances individually and then, using instance labels, infer the label of the bag. These methods are suitable for witness identification. However, classifying bags differs from classifying individual instances. For example, under the standard MIL assumption that a positive bag contains at least one positive instance, when classifying a bag, once a positive instance has been identified, false negatives have no impact. Therefore, the best bag classifier is not necessarily the best instance classifier [17]. This section describes the witness identification strategy of several instance-based MIL methods.

The simplest approach, which is not a MIL method *per se*, is to consider that the label of each instance corresponds to the label of the bag it belongs to, and train a regular supervised classifier. The negative instances in positive bags add noise to the optimization process. If the proportion of noise is low, this method performs relatively well, but performances rapidly decrease when the WR is low.

One of the first MIL methods, APR [2] searches for a hyper-rectangle in feature space containing mostly instances from positive bags, and as few as possible instances from negative bags. The instances the hyper-rectangle encompasses are considered to be witnesses. While this method is successful in some situations, it has problems dealing with multimodal positive data distributions.

Two of the first MIL methods based on SVMs, mi-SVM and MI-SVM were proposed in the same paper [3]. Both methods intrinsically perform witness identification, but differ in the strategy used to discover witnesses. In mi-SVM, the margin is maximized jointly over the discriminant function and individual instance label assignations of the complete data set. At first, a label is assigned to each instance, and an SVM is trained based on the instance label attribution. Instances are then reclassified using the newly trained SVM. The resulting labels are then assigned to each instance and the SVM is retrained. This procedure is repeated until the labels are stable. The witnesses are the instances with a positive label. MI-SVM uses the same iterative procedure, except that positive bags are represented by the single most positive instance in the bag. Because it selects only one instance in each bag, this method has problems dealing with bags containing positive instances from more than one concept.

Instead of looking for witnesses directly, Maron and Lorenzo-Pérez proposed a measure called diverse density (DD) [18]. This measures the probability that a given point in feature space belongs to the positive class. It depends on the proportion of instances from positive and negative bags in the neighborhood. The highest point of the DD function corresponds to the positive concept from which are generated the witnesses, and instances are classified based on their proximity to this point. Later, in EM-DD [19], the Expectation-Maximization algorithm was used to locate the maximum of the DD function. Because these methods seek a single maximum point, they assume that positive instances come from a single compact cluster in feature space, which limits their applicability to many problems. It has also been pointed out that EM-DD performance decreases when the number of noisy features increases [19]. DD and SVM are combined in DD-SVM [20]. Local maxima of the DD function are selected and used as prototypes. The distances between the prototypes and the instances in bags are used as feature vectors, which are classified by an SVM. MILES [5] uses the same kind of distance-based embedding except that the prototypes are replaced by instances selected from the data set using a 1-norm SVM. The authors provided a way to identify witness based on each instance contribution to the bag label.

Some methods were proposed specifically to locate regions of interest (ROI) in images for CBIR. For example, CkNN-ROI [15] classifies bags using the Hausdorff distance and the reference and citations scheme of Citation-kNN [11]. Once a bag is deemed positive, each instance it contains is treated as a bag, and is classified individually. The instances classified as positive are the witnesses. KI-SVM [4] also locates ROI by finding the key instance (i.e. witness) in bags using multiple kernel learning. The program is constrained to correctly classify each instance in negative bags. In the MKL formulation, each possible instance label assignation in positive bags corresponds to a kernel. The algorithm seeks a combination of kernels which produces a correct label assignment in the data set. During its optimization, the constraints are satisfied if the bags are correctly labeled, and thus, if the positive bags contain more than one witness from different concepts, some witnesses can be ignored.

Most of the methods are less effective when the WR is low, or when the data sets contain two or more positive concepts. The proposed algorithm, RSWI (see next Section), consistently provides a high level of performance; it is robust to a large range of WR and allows to learn from multi-concept distributions.

III. Random Subspace Witness Identification

In this paper a new method called RSWI is proposed. It identifies witnesses by analyzing the neighborhood composition of each instance. The neighborhoods are defined by clusters in multiple random subspaces. The method is related to DD in the sense that this is a measure of the likelihood that an instance is positive, but instead of locations in feature space, a score is given to instances. An advantage of RSWI is that there is no search for a global maximum, which makes the method robust to multimodal distributions. Moreover, RSWI performs a series of simple tasks which are computationally efficient. (see Figure 1).

In MIL problems $\mathcal{B} = \left\{B^1, ..., B^Z\right\}$ is a set of Z bags, each corresponding to a label $L^i \in \{-1, +1\}$. Each bag

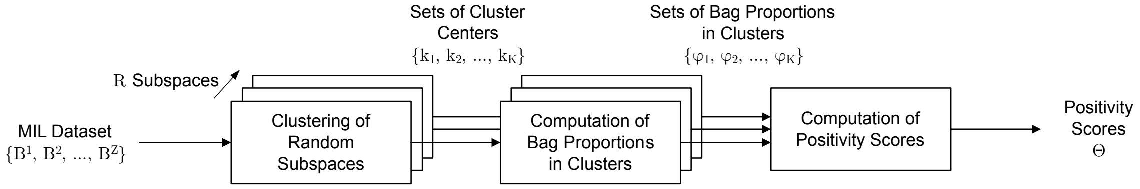

Sets of Cluster Centers
{k1, k2, ..., kK}
Sets of Bag Proportions in Clusters
{φ1, φ2, ..., φK}
R Subspaces
MIL Dataset
{B1, B2, ..., BZ}
Clustering of Random Subspaces
Computation of Bag Proportions in Clusters
Computation of Positivity Scores
Positivity Scores
Θ

Fig. 1. Block diagram for positivity scores computation.

contains N^i couples composed of a feature vector and its associated label: $B^i = \left\{(\mathbf{x}_1^i, y_1^i), ..., (\mathbf{x}_{N^i}^i, y_{N^i}^i)\right\}$ where $\mathbf{x}_j^i = (x_{j1}^i, ..., x_{jd}^i) \in \mathbb{R}^d$. The labels y_j^i of each individual instance are unknown in positive bags, but are assumed to be negative in negative bags. Following the standard MIL assumption [1], there is at least one positive instance per positive bag.

With RSWI, instances are identified based on a *positivity score* computed as follows: At first, subspaces $\mathcal{P}$ are created by randomly selecting p features from the complete set of d features. Every instance $\mathbf{x}$ in the data set is projected in the p-dimensional subspaces. Next, the data in each subspace is clustered. Here, a hard assignment method (e.g. k-means), is assumed, but any clustering algorithm could be used. Each subspace captures a different relation between instances resulting in different clusterings. The second step consists in computing the proportion φ_n of instances belonging to positive bags in each cluster k_n:

$$\varphi_n = \frac{\sum_{\forall \mathbf{x}} c(\mathbf{x}^i, n)}{|\mathcal{K}_n|} \in [0..1], \tag{1}$$

where $n = 1, 2, ..., K$, and

$$c(\mathbf{x}^i, n) = \begin{cases} 1, & \text{if } \mathbf{x}^i \in \mathcal{K}_n \text{ and } L^i = +1; \\ 0, & \text{otherwise.} \end{cases}. \tag{2}$$

In these equations, $\mathcal{K}_n$ represents the set of instances belonging to cluster k_n. The size of this set is given by $|\mathcal{K}_n|$.

These two steps (projection into a random subspace and clustering), are repeated R times. The third and last step is the computation of the instances positivity score $\theta(\mathbf{x})$. This score is the mean of all the positive bag proportion $\varphi_n(r)$ of the clusters it was assigned to:

$$\theta(\mathbf{x}) = \frac{1}{R} \sum_{r=1}^{R} \sum_{n=1}^{K} \varphi_n(r) \cdot d(\mathbf{x}, n, r), \tag{3}$$

where

$$d(\mathbf{x}, n, r) = \begin{cases} 1, & \text{if } \mathbf{x} \in \mathcal{K}_n \text{ at repetion } r; \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

These positivity scores give an indication of the likelihood that an instance is a witness. The label for $\mathbf{x}$ is given by:

$$y = \begin{cases} +1, & \text{if } \theta(\mathbf{x}) > \alpha; \\ -1, & \text{otherwise,} \end{cases} \tag{5}$$

where α is the decision threshold. If labeled instances are available, this threshold should be optimized based on the desired performance measure. However, in most MIL problems, instances labels are unavailable. In that case, the threshold can be set by making sure at least one instance is classified as a witness in each positive bag:

$$\alpha = \min_{B_i \in \mathcal{B}^+} \left\{ \max_{\mathbf{x} \in B_i} \theta(\mathbf{x}) \right\} \tag{6}$$

where $\mathcal{B}^+$ is the set containing only the positive bags. Following (6), there will be at least one bag containing only one witness, but the other bags may contain any number.

Because RSWI is a local measure of positivity, it allows to identify witnesses in different regions of the feature space, making the algorithm robust to multimodal distributions. Also, since this measure is relative to all instances in the data set, witnesses can be identified reliably regardless of the WR.

IV. Experimental Methodology

In many MIL papers, the accuracy is used as a performance metric. While reasonable when evaluating bag classification, it may be misleading in the context of instance classification, where class data is unbalanced. For example, in a data set where the WR is 20% and there are an equal number of negative and positive bags, predicting only negative instances would achieve an accuracy of 90%. This is why the area under the receiver operating characteristic curve (AUC) and the area under the precision-recall curve (AUPRC) will be used in this paper as primary comparison metrics. To measure the ability of the algorithm to select a decision threshold, the F_1-scores will also be reported. The F_1-score is the harmonic mean between precision and recall. Since the negative bags are assumed to contain only negative instances, they are not relevant for the comparison on witness identification, and thus, are ignored when measuring performance. For data sets generated several times, both the average results and standard deviations are reported.

Some algorithms have parameters that need to be optimized on the data. This is done via grid-search using 5-fold cross-validation on the entire data set. Since the instance labels are unknown, the performance of each configuration is evaluated using bag-level AUC. For the RSWI algorithm, two parameters were optimized. The dimensionality of the random subspaces ranged from 20% to 50% of the complete feature space dimensionality. The number of clusters used in k-means ranges from 30 to 120 with steps of 30. In all experiments, 2000 random subspaces were generated, this has proved to provide stable results in previous experiments. Fewer subspaces can be used especially with low-dimensional data sets, but since the method is computationally inexpensive, this parameter was not optimized. For all methods involving SVM, the regularization parameter (C) ranged from 0.1 to 10000 and the spread of the RBF kernel (γ), from 0.01 to 1000.

A. Reference Methods

SI-SVM: SI-SVM is an SVM trained using the labels assigned to bags as instance labels. It gives an indication on the pertinence of using MIL methods instead of regular supervised algorithms in a problem. The LIBSVM [21] implementation has been used.

CkNN-ROI: This method was selected because it was proposed for the identification of regions of interest (i.e. witness) in CIBR tasks. The method was implemented based on the details provided in the paper and the CkNN implementation provided on Zhou's website. The number of citers and references, ranging from 1 to 9 are chosen by grid-search cross-validation.

MI-SVM & mi-SVM: The two algorithms were implemented as described in the original paper [3]. The LIBSVM [21] implementation has been used, and the parameters were optimized at each algorithm iteration.

EM-DD: The method has been selected as a reference method because it is the algorithm with the closest objective to the proposed method. The implementation provided with the MIL toolbox was used [22]. The algorithm was reinitialized 20 times, starting at the position of a random instance belonging to a positive bag. Only the result from the best run is used.

MILES: This method has been selected because it performs well on benchmark data sets and because the authors provided a way to use their algorithm for instance naming. The implementation provided with the MIL toolbox [22] has been used.

KI-SVM: This method has been selected because it has been designed to find the key instance (i.e. witness) in bags. Since the bag-level version is a simplification of the instance-level version, only the instance-level version was used in this paper. The implementation provided by the authors on Zhou's website was used in the experiments.

B. Data Sets

Most existing MIL data sets do not provide annotation of individual instances. Therefore the Letters and Mammograms MIL data sets described below have been created using real-world data from existing data sets to evaluate MIL algorithms on the witness identification task.

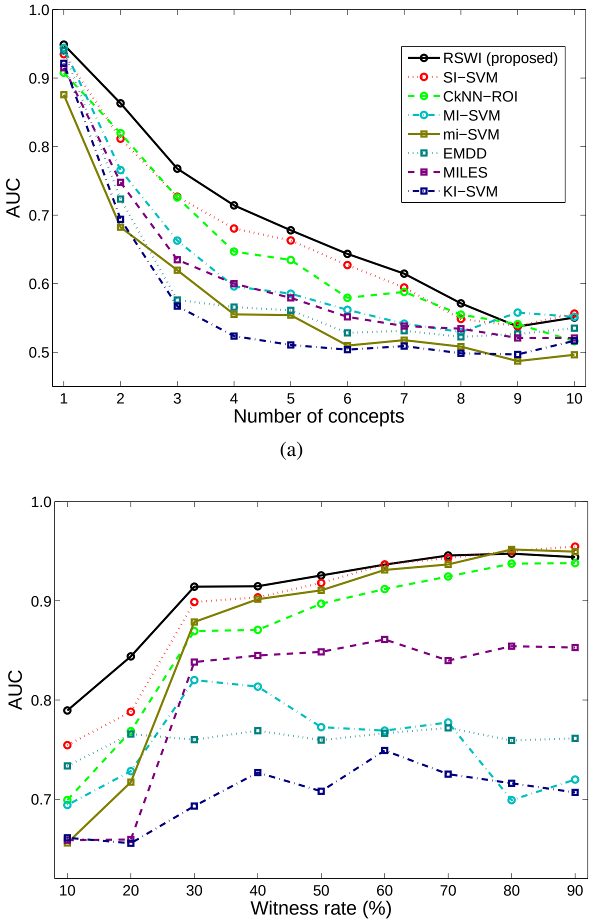
Letters: This data set is created using the Letter Recognition data set introduced in [23]. It contains a total of 20k instances of the 26 letters in the English alphabet. Each letter is encoded by a 16-dimensional feature vector. The reader is referred to the original paper for more details. A MIL version of the data set is created by grouping letters in bags. This allows control over WR and the number of positive concepts, which in this context, correspond to the different letters. A first collection of data sets is created by varying the number of positive concepts from 1 to 10. Each time a data set is generated, random letters are designated to be positive concepts, and all others are assigned to negative concepts. All bags contain 10 instances, and positive bags contain 2 instances from randomly selected from the positive concept. A second collection of data sets is generated to assess the effects of WR. The positive class is composed of 3 randomly selected concepts. Each bag contains 10 instances, and the number of witnesses in positive bags is determined by the WR. All data sets contain 100 positive and 100 negative bags. For each configuration, 10 different data sets are generated.

Birds: The birds data set was introduced in [24]. In this data set, each bag corresponds to a 10 seconds recording of bird songs from one or more species. The recording is temporally segmented, and each part corresponds to a particular bird, or to background noises. These segments are the instances, each of represented by 38 features. Details on the features are given in the original paper. There are 13 types of bird in the data set. If one specie at a time is considered as the positive class, 13 MIL problems can be generated from this data set. Due to space constraints, only the results for the species providing the least and the most number of witnesses were reported. The entire data set contains a total 10232 instances, of which 32 belong to the hermit thrush and 1280 to the Hammond's flycatcher. The difficulty for MIL is that the WR is low and is not constant across positive bags.

Mammograms: This data set is created from the images contained in mini-MIAS database of mammograms [25]. The database contains images of healthy patients, as well as patients exhibiting 1 of the 6 classes of abnormalities. For each abnormality, an image patch is extracted using the location annotations provided with the data set. These patches are positive instances, and negative instances are patches of various sizes extracted from tissue regions not intersecting with abnormalities regions, or from tissue regions belonging to healthy patients. Each patient is represented by a bag containing 10 patches. Because negative patches are extracted randomly, 5 versions of the data set are generated. The data set contains a total of 326 subjects, among which there are 117 subjects presenting abnormalities. Features are extracted from each patch. Similarly to [7], the feature vector contains the mean and standard deviation and a normalized 12-bin frequency histogram of the pixel intensities contained in the patch. This representation is augmented with the mean local binary pattern (LBP) extracted from a 13×13 pixel grid, and with the mean of densely extracted SIFT descriptors. Finally, the 5 Haralick features used in [26] are also used. The resulting 220-dimensional vectors are reduced to 100-dimensional vectors using PCA. The difficulty for MIL is that the WR is low and there are 6 concepts in the positive class.

V. Results

Fig. 2 shows the mean AUC of proposed and reference methods vs. the number of positive concepts and WR on the Letters data set. The AUPRC and F_1-score were not reported due to space constraints, and because they did not provide contrasting information to the AUC curve. In the number of concepts experiments, the performance of all algorithms decreases as the problem complexity increases. However,



(a)

(b)

Fig. 2. Performance of MIL algorithms on the Letters data set.

TABLE I
PERFORMANCE ON THE MAMMOGRAMS AND BIRDS MIL DATA SET.

Method	AUC (×100)	AUPRC (×100)	F$_1$-score (%)
Mammography (WR = 10%)			
SI-SVM	53.1 (5.8)	11.3 (2.3)	18.3 (0.2)
CkNN-ROI	56.7 (2.1)	14.6 (3.7)	20.2 (1.1)
MI-SVM	**69.3 (9.0)**	**26.6 (11.9)**	**26.4 (11.0)**
mi-SVM	53.4 (7.2)	13.7 (5.8)	18.8 (0.8)
EM-DD	55.6 (7.8)	13.6 (2.7)	8.8 (4.3)
MILES	65.5 (2.3)	24.0 (4.5)	23.8 (1.3)
KI-SVM	55.1 (10.1)	14.0 (6.3)	1.3 (2.1)
Proposed (RSWI)	67.4 (1.6)	26.2 (1.6)	24.1 (2.1)
Hermit Thrush (32/10232 witnesses)			
SI-SVM	61.1	12.4	8.6
CkNN-ROI	59.5	14.6	0.0
MI-SVM	59.2	16.4	5.2
mi-SVM	**70.7**	15.4	8.7
EM-DD	44.8	0.0	0.0
MILES	52.4	17.2	12.2
KI-SVM	37.1	7.3	0.0
Proposed (RSWI)	68.3	**20.5**	**29.1**
Hammond's Flycatcher (1280/10232 witnesses)			
SI-SVM	87.9	97.1	89.9
CkNN-ROI	89.4	97.6	89.6
MI-SVM	84.6	96.6	17.5
mi-SVM	89.0	97.6	**90.0**
EM-DD	89.2	97.8	58.9
MILES	74.8	93.7	55.0
KI-SVM	86.4	96.8	60.8
Proposed (RSWI)	**91.0**	**98.2**	86.6

three methods, RSWI, CkNN and SI-SVM, are affected to a lesser extent. Both RSWI and CkNN-ROI are non-parametric methods, in which instances are classified based on bag distribution in their neighborhood. These local approaches provide robustness to distribution shape when compared to methods where an optimization process is performed using a global objective on all the data set. While CkNN-ROI is robust to the number of clusters, it is affected by low WR. Instances are labeled positive if they are close to any of the instances of a positive bag. If positive bags contain a large proportion of negative instances, it is more likely that negative test instances are found to be close to positive bags, which results in a high false positive rate. RSWI is affected by low WR to a lesser extent than all of the other methods. This is because witnesses are identified by comparing scores representing the proportion of instances from positive bags in their neighborhood. Even if this score is high in negative regions, it will still be lower than in positive regions of the feature space.

SI-SVM dominates all other SVM-based methods and EM-DD. It has been found in the past that in some application, SI-SVM may perform as well, or better the MIL algorithms [27]. With SI-SVM, the problem is reduced to classification with a one-sided class noise. This reasonably applies to this task because the positive instances are organized in a small number of compact clusters, while negative instances are well distributed in feature-space in a greater number of clusters. SI-SVM is the first iteration of mi-SVM. This indicates that the iterative optimization procedure of relabeling and training slowly converts positive regions of the feature space into negative regions. This happens when the number of positive instances is limited and distributed in many clusters. These positive regions become scanty, and thus, more susceptible to misclassification. However, as observed in Fig. 2 (b), when the WR increases mi-SVM performs comparably to SI-SVM.

As for KI-SVM and MI-SVM, during optimization, witnesses are selected under the constraint of bag classification accuracy. Only one instance per bag is selected, which is enough under the standard MIL assumption to achieve high levels of bag classification accuracy. In a witness identification task, however, the goal is not to identify at least one witness, but all witnesses. If all bags contain positive instances from two or more concepts, the instances from one concept are predominantly selected, and thus, the others are ignored, which leads to poor performances. A similar argument can be made for MILES, which constructs a bag representation from an instance selection process governed by bag-level classification accuracy. EM-DD performance also declines when there is more than one concept. This is expected, since the algorithm searches for a single maximum of the DD function corresponding to the dominating concept. All other concepts are ignored.

The performance of the proposed and reference techniques on the Mammograms and Birds data sets is shown in Table I. The Mammograms data set has a low WR (10%) and is composed of multiple positive concepts corresponding to the 6 abnormality classes. MI-SVM is the best-performing algorithm despite the previous observation that this algorithm is affected by the presence of multiple concepts. In the Letters case, there is more than one witness per bag, although the algorithm selects only one during optimization. In the Mammograms data set, however, there is only one witness per bag, and thus selecting only one instance does not affect MI-SVM performance. The results obtained by RSWI are slightly lower to those obtained with MI-SVM. However, the results standard deviations indicate that RSWI achieves a high level of performance more consistently across all versions of the data set, which is a desirable property in practice.

The experiments on the Birds data set show the robustness of the proposed method to low WR. In the case of the Hermit Thrush, the witnesses represent only 0.3% of all instances. In such extreme conditions, many methods fail. For example, CkNN-ROI, EM-DD and KI-SVM cannot detect any of the witnesses, and thus, obtain a F_1-score of 0. SI-SVM and mi-SVM obtain appreciable results in terms of AUC but did not perform well in terms of F_1-score. Results suggest that both methods struggled to find an optimal classification threshold, which is the offset of the SVM hyper-plane. Both methods assume that all instances in positive bags are positive, which causes the SVM to include incorrectly labeled negatives in the positive instance region. When the number of witnesses in the data set increases, as in the Hammond Flycatcher case, most algorithms perform comparably. However MI-SVM and KI-SVM do not achieve the performance level of their counterparts because both algorithms assume there is only one witness per bag which is not the case in this data set.

VI. Conclusion

This paper presents a new MIL method for witness identification called RSWI. The proposed method achieves a high level of performance in all 3 tested applications, and demonstrated its applicability to problems with low WR and multiple positive concepts. The method is compared to 7 reference methods and obtains the best overall performance and consistently achieves first or second rank, while other methods perform unevenly across applications.

Future research will include methods to find a better classification threshold for the proposed and the reference methods. In addition, usability of RSWI as a component of a MIL algorithm should be explored.

References

[1] J. Amores, "Multiple Instance Classification: Review, Taxonomy and Comparative Study," *Artif. Intell.*, vol. 201, pp. 81–105, Aug. 2013.

[2] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, "Solving the Multiple Instance Problem with Axis-parallel Rectangles," *Artif. Intell.*, vol. 89, no. 1-2, pp. 31–71, Jan. 1997.

[3] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support Vector Machines for Multiple-Instance Learning," in *Advances in Neural Infor. Process. Syst. 15*. MIT Press, 2002, pp. 561–568.

[4] Y.-F. Li, J. T. Kwok, I. W. Tsang, and Z.-H. Zhou, "A Convex Method for Locating Regions of Interest with Multi-instance Learning," in *European Conf. on Mach. Learning and Knowl. Discovery in Databases: Part II*. Berlin, Heidelberg: Springer, 2009, pp. 15–30.

[5] Y. Chen, J. Bi, and J. Z. Wang, "MILES: Multiple-Instance Learning via Embedded Instance Selection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 12, pp. 1931–1947, Dec. 2006.

[6] G. Quellec, M. Lamard, M. Cozic, G. Coatrieux, and G. Cazuguel, "Multiple-Instance Learning for Anomaly Detection in Digital Mammography," *IEEE Trans. Med. Imag.*, vol. PP, 2016.

[7] M. Kandemir, A. Feuchtinger, A. Walch, and F. A. Hamprecht, "Digital pathology: Multiple instance learning can detect Barrett's cancer," in *Int. Symp. on Biomedical Imaging*, 2014, pp. 1348–1351.

[8] J. Melendez, B. van Ginneken, P. Maduskar, R. Philipsen, H. Ayles, and C. Sanchez, "On combining multiple-instance learning and active learning for computer-aided detection of tuberculosis," *IEEE Trans. Med. Imag.*, vol. PP, 2015.

[9] D. Palachanis, "Using the Multiple Instance Learning framework to address differential regulation," Master, Delft University of Technology, 2014.

[10] G. Mohammadi and A. Vinciarelli, "Automatic Personality Perception: Prediction of Trait Attribution Based on Prosodic Features," *Affective Computing, IEEE Trans. on*, vol. 3, no. 3, pp. 273–284, jul 2012.

[11] J. Wang and J.-D. Zucker, "Solving the Multiple-Instance Problem: A Lazy Learning Approach," in *Int. Conf. on Mach. Learning*. San Francisco, USA: ACM, 2000, pp. 1119–1126.

[12] V. Cheplygina, D. M. J. Tax, and M. Loog, "Dissimilarity-Based Ensembles for Multiple Instance Learning," *IEEE Trans. Neural Netw. Learn. Syst.*, pp. 1–13, 2015.

[13] R. C. Bunescu and R. J. Mooney, "Multiple Instance Learning for Sparse Positive Bags," in *Int. Conf. on Mach. Learn.* New York, USA: ACM, 2007, pp. 105–112.

[14] Z.-H. Zhou and M.-L. Zhang, "Solving Multi-instance Problems with Classifier Ensemble Based on Constructive Clustering," *Knowl. Infor. Syst.*, vol. 11, no. 2, pp. 155–170, 2007.

[15] Z.-H. Zhou, X.-B. Xue, and Y. Jiang, "Locating Regions of Interest in CBIR with Multi-instance Learning Techniques," in *Australian Joint Conf. on Artificial Intelligence*. Springer, 2005, pp. 92–101.

[16] M.-A. Carbonneau, E. Granger, A. J. Raymond, and G. Gagnon, "Robust multiple-instance learning ensembles using random subspace instance selection," *Pattern Recognition*, vol. 58, pp. 83–99, 2016.

[17] G. Vanwinckelen, V. do O, D. Fierens, and H. Blockeel, "Instance-level accuracy versus bag-level accuracy in multi-instance learning," *Data Mining and Knowl. Discovery*, pp. 1–29, 2015.

[18] O. Maron and T. Lozano-Pérez, "A Framework for Multiple-Instance Learning," in *Advances in Neural Infor. Process. Syst.* Cambridge, USA: MIT Press, 1998, pp. 570–576.

[19] Q. Zhang and S. A. Goldman, "EM-DD : An Improved Multiple-Instance Learning Technique," in *Advances in Neural Infor. Process. Syst.* MIT Press, 2001, pp. 1073–1080.

[20] Y. Chen and J. Z. Wang, "Image Categorization by Learning and Reasoning with Regions," *Journal of Mach. Learning Research*, vol. 5, pp. 913–939, Dec. 2004.

[21] C.-C. Chang and C.-J. Lin, "LIBSVM: A Library for Support Vector Machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 1–27, May 2011.

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

[23] P. Frey and D. Slate, "Letter recognition using holland-style adaptive classifiers," *Mach. Learn.*, vol. 6, no. 2, pp. 161–182, Mar. 1991.

[24] F. Briggs, X. Z. Fern, and R. Raich, "Rank-loss support instance machines for MIML instance annotation," in *Int. Conf. on Knowl. Discovery and Data Mining*. ACM, 2012, pp. 534–542.

[25] J. Suckling, J. Parker, D. Dance, S. Astley, I. Hutt, C. Boggis, I. Ricketts, E. Stamatakis, N. Cerneaz, S. Kok, and Others, "The mammographic image analysis society digital mammogram database," in *Exerpta Medica. Int. Congr. Series*, vol. 1069, 1994, pp. 375–378.

[26] N. R. Mudigonda, R. M. Rangayyan, and J. E. L. Desautels, "Gradient and texture analysis for the classification of mammographic masses," *IEEE Trans. Med. Imag.*, vol. 19, no. 10, pp. 1032–1043, 2000.

[27] S. Ray and M. Craven, "Supervised Versus Multiple Instance Learning: An Empirical Comparison," in *Int. Conf. on Mach. Learning*. New York, USA: ACM, 2005, pp. 697–704.