ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 7, Issue 8, August 2018

Survey of Multi Instance learning Algorithms

M.Kavitha[1], Jasmin Thomas[2]

Assistant Professor, PG and Research Department Of Computer Science,

Tirupur Kumaran College for Women, Tirupur, India[1]

Research Scholar, PG and Research Department Of Computer Science,

Tirupur Kumaran College for Women, Tirupur, India[2]

Abstract: In multi-instance learning, the training set comprises labelled bags that are composed of unlabeled instances, and the task is to predict the labels of unseen bags. The Multiple instance learning (MIL) is a form of weakly supervised learning where training instances are arranged in sets, called bags, and a label is provided for the entire bag. The supervised learning where training instances are arranged in sets, called bags, and a label is provided for the entire bag. This formulation is gaining interest because it naturally fits various problems and allows to leverage weakly labelled data. However, learning from bags raises important challenges that are unique to MIL. This paper provides a complete survey of the characteristics which define and distinguish the types of MIL problems. Until now, these problem characteristics have not been properly identified and described. There are some types learning generally in ML algorithms like, the types of data distribution, the ambiguity of instance labels, and the task to be performed. Some important issues to be addressed in this paper. Finally delivers insight on how the problem characteristics affect MIL algorithms, recommendations for future benchmarking.

Keywords: Multiple instance learning, weakly labelled data, ambiguity of instance labels

I. INTRODUCTION

During the investigation of drug activity prediction, the Multi-Instance Learning (MIL) framework was formally proposed and naturally applied to this problem. In contrast to traditional supervised learning, MIL receives a set of bags labelled positive or negative, rather than receiving a set of instances which have labels. In addition, instances in the MIL bags have no label information. The task of MIL is to train a classifier that labels new bags, and MIL has already been widely applied in diverse applications, e.g., image categorization, text categorization, face detection, computer-aided medical diagnosis, web mining, etc. The many effective MIL algorithms have been developed. These algorithms achieve decent accuracy rates in different MIL applications, which might partly attribute to the fact that objects are represented as bags in MIL, which can naturally encode the original objects. However, directly processing and classifying the complicated bag representation means that the complexity of MIL's hypothesis space also becomes much larger. This fact leads to an undesired outcome: most existing MIL algorithms are usually time-consuming and incapable of handling large scale MIL problems. The real-world applications of MIL, however, consistently request scalable multi-instance learning algorithms to handle millions of complex objects or examples (e.g., images, genes, etc). During the past years, learning from examples is one of the most flourishing areas in machine learning. According to the ambiguity of training data, research in this area can be roughly categorized into three learning frameworks, i.e. supervised learning, unsupervised learning, and reinforcement learning. Supervised learning attempts to learn a concept for correctly labelling unseen instances, where the training instances are with known labels and therefore the ambiguity is the minimum. Unsupervised learning attempts to learn the structure of the underlying sources of instances, where the training instances are without known labels and therefore the ambiguity is the maximum. Reinforcement learning attempts to learn a mapping from states to actions, where the instances are with no labels but with delayed rewards that could be viewed as delayed labels, therefore the ambiguity is between that of supervised learning and unsupervised learning.

Different to supervised learning where all training instances are with known labels, in multi-instance learning the labels of the training instances are unknown; different to unsupervised learning where all training instances are without known labels, in multi-instance learning the labels of the training bags are known; different to reinforcement learning where the labels of the training instances are delayed, in multi-instance learning there is no any delay. Since multi-instance problems extensively exist but are unique to those addressed by previous learning frameworks, multi-instance learning was regarded as a new learning framework, and has attracted much attention of the machine learning community. The term multi-instance learning was coined by Dietterich et al. when they were investigating the problem of drug activity prediction. In multi-instance learning the training set is composed of many bags each contains many

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 7, Issue 8, August 2018

instances. A bag is positively labelled if it contains at least one positive instance; otherwise it is labelled as a negative bag. The task is to learn some concept from the training set for correctly labelling unseen bags.

II. MIL FRAMEWORK

MIL classification is not limited to assigning a single label to instances or bags. Assigning multiple labels to bags is particularly relevant considering that they can contain instances representing different concepts.

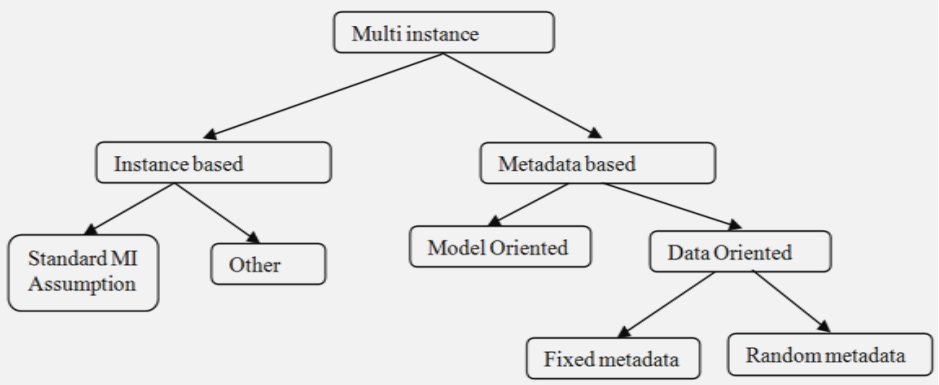

Fig 1: MIL Framework

A. Instance-based algorithms: The term "instance-based" denotes that the algorithm attempts to find a set of representative instances based on an MI assumption and classify future bags from these representatives.

B. Metadata-based algorithms: The metadata-based algorithms make no assumptions about the relationship between instances and bag labels, and instead try to extract instance-independent information (or metadata) about the bags in order to learn the concept.

III. RELATED WORK

Because many problems can be formulated as MIL, there is a plethora of MIL algorithms in the literature. However, there is only a handful of general MIL studies and surveys. This section summarizes and interprets the broad conclusions from these general MIL papers. R. Hong et., Image annotation by multiple-instance learning with discriminative feature mapping and selection, propose a MIL method with discriminative feature mapping and feature selection, aiming at solving this problem. Our method is able to explore both the positive and negative concept correlations. It can also select the effective features from a large and diverse set of low-level features for each concept under MIL settings. MIL approach consists of two stages. In the first stage, each bag is mapped to a new feature vector in bag-level feature space, turning the MIL problem to a standard single instance learning problem. The mapped bag-level features are often high-dimensional and with much noise. Therefore, in the second step, an AdaBoost procedure is implemented to select the bag-level features and build the final classifier. The disadvantage of our proposed method, i.e., the annotation was based on a single decision and not based on multiple possibilities decision.

H. Yuan et., al., Hierarchical sampling for multi-instance ensemble learning, propose a Hierarchical Sampling-based Multi-Instance ensemble Learning (HSMILE) method. Due to the unique multi-instance learning nature, a positive bag contains at least one positive instance whereas samples (instance and sample are interchangeable terms in this paper) in a negative bag are all negative, simply applying bootstrap sampling to individual bags may severely damage a positive bag because a sampled positive bag may not contain any positive sample at all. The disadvantage of this paper is without considering bag constraint can achieve maximum diversity, but may end up forming a positive bag containing negative samples only.

Q. Wang et., al., Saliency detection by multiple instance learning, the obtained saliency map is also inconsistent with many properties of human behaviour. In order to overcome the challenges of inability and inconsistency, this paper presents a framework based on multiple-instance learning. Low-, mid-, and high-level features are incorporated in the detection procedure, and the learning ability enables it robust to noise. The disadvantage of this proposed system is saliency maps of several algorithms suffer from low resolution and light object boundaries but fail to detect the whole target region.

J. Wu et., al., Bag constrained structure pattern mining for multi-graph classification, multi-graph learning is to build a learning model from a number of labelled training bags to predict previously unseen test bags with maximum accuracy. This problem setting is essentially different from existing multi-instance learning (MIL), where instances in MIL share

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 7, Issue 8, August 2018

well defined feature values, but no features are available to represent graphs in multi-graph bags. To solve the problem, we propose a Multi-Graph Feature based Learning (gMGFL) algorithm that explores and selects a set of discriminative sub graphs as features to transfer each bag into a single instance, with the bag label being propagated to the transferred instance. This method has some disadvantages like graph similarity is assessed based on the global graph structures, such as paths or walks, between graphs. So, it is not clear which substructures (or which parts of the graphs) contribute to the most to the similarity assessment.

J. Amores, Multiple instance classification: Review, taxonomy and comparative study, Multiple Instance Learning (MIL) has become an important topic in the pattern recognition community, and many solutions to this problem have been proposed until now. Despite this fact, there is a lack of comparative studies that shed light into the characteristics and behaviour of the different methods. In this work we provide such an analysis focused on the classification task (i.e., leaving out other learning tasks such as regression). In order to perform our study, we implemented fourteen methods grouped into three different families. We analyse the performance of the approaches across a variety of well-known databases, and we also study their behaviour in synthetic scenarios in order to highlight their characteristics. SMI assumption do not seem to have an impact on the performance.

Z. Fu et., al., Milis: Multiple instance learning with instance selection, propose MILIS, a novel MIL algorithm based on adaptive instance selection. We do this in an alternating optimisation framework by intertwining the steps of instance selection and classifier learning in an iterative manner which is guaranteed to converge. Initial instance selection is achieved by a simple yet effective kernel density estimator on the negative instances. This proposed system has disadvantage This is not a condition of the algorithm since any number of IPs can be chosen from each bag for feature mapping. Also, note that, for negative bags, we have chosen the most negative instance as the initial IP.

J. Foulds and E. Frank, revisiting multiple-instance learning via embedded instance selection, our results show that boosted decision stumps can in some cases provide better classification accuracy than the 1-norm SVM as a base learner for MILES. Although MILES provides competitive performance when compared to other MI learners, we identify simpler propositionalizing methods that require shorter training times while retaining MILES' strong classification performance on the datasets we tested. The disadvantage of our proposed method distance-based method does not efficient one.

J. Wu et., al., Multi-graph learning with positive and unlabeled bags, to solve the challenge, we propose a puMGL learning framework which relies on two iteratively combined processes for multigraph learning: (1) deriving features to represent graphs for learning; and (2) deriving discriminative models with only positive and unlabeled graph bags. For the former, we derive a subgraph scoring criterion to select a set of informative subgraphs to convert each graph into a feature space. To handle unlabeled bags, we assign a weight value to each bag and use the adjusted weight values to select most promising unlabeled bags as negative bags. A margin graph pool (MGP), which contains some representative graphs from positive bags and identified negative bags, is used for selecting subgraphs and training graph classifiers. The iterative subgraph scoring, bag weight updating, and MGP based graph classification forms a closed loop to find optimal subgraphs and most suitable unlabeled bags for multi-graph learning. The disadvantage os this proposed system is unchecked images, they may not contain users' retrieval concepts (i.e. negative bags) or users simply overlook the images. In this case, there is no negative bag but only positive and unlabeled bags are available.
D. Nguyen et., al., Multiple-instance learning (MIL) is a supervised learning technique that addresses the problem of classifying bags of instances instead of single instances. In this paper, we introduce a rule-based MIL algorithm, called mi-DS, and compare it with 21 existing MIL algorithms on 26 commonly used data sets. The results show that mi-DS performs on par with or better than several well-known algorithms and generates models characterized by balanced values of precision and recall. The new rule-based MIL algorithm, called mi-DS, and compared it with 21 MIL algorithms on 26 data sets that ranged from numerical, through text, to images. The results indicated that, although there did not exist one generally best performing algorithm on all data sets, the mi-DS performed on average quite well and had desirable characteristics that distinguished it from other algorithms. First, it showed very good predictive accuracy on most data sets as measured by the average accuracy rank. In particular, it did well on text and pre-processed image data sets. Second, the differences between mi-DS and nine other algorithms were statistically significant for the six of them. Third, it performed quite well on data with missing values. Fourth, it is faster than most of the algorithms it was compared with. Importantly, the approach used in the mi-DS can be used as a generic framework for modifying other rule-based algorithms so that they can be used for solving MIL problems.

V. Cheplygina et., al., Multiple instance learning with bag dissimilarities, Multiple Instance Learning (MIL) is concerned with learning from sets (bags) of objects (instances), where the individual instance labels are ambiguous. In this setting, supervised learning cannot be applied directly. Often, specialized MIL methods learn by making additional assumptions about the relationship of the bag labels and instance labels. Such assumptions may fit a particular dataset,

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 7, Issue 8, August 2018

but do not generalize to the whole range of MIL problems. Other MIL methods shift the focus of assumptions from the labels to the overall (DIS)similarity of bags, and therefore learn from bags directly. We propose to represent each bag by a vector of its dissimilarities to other bags in the training set, and treat these dissimilarities as a feature representation. We show several alternatives to define a dissimilarity between bags and discuss which definitions are more suitable for particular MIL problems. It would be interesting to investigate the exact trade-off of these two choices. Overall, we believe the proposed approach is a flexible, powerful and intuitive way to do MIL, and that combined, these qualities make it an attractive method for domains where data might be naturally grouped in bags, but MIL is not yet being used.

IV. MIL WITH DISCRIMINATIVE BAG MAPPING

In multi-instance learning, a bag B_i consists of a set of instances, with $x_{i,j}$ denoting the j^{th} instance in B_i. The class label of Bi is denoted by $y_i = \mathcal{Y}$, with $\mathcal{Y} = \{-1, +1\}$. The collection of all bags can be denoted by B. In our proposed bag mapping framework, each bag Bi will be transformed to B i , a single instance in a new feature space by using the discriminative instance pool (DIP, denoted by P). Given B with n bags, and the instance set X collected from all bags in B, Our objective is to find a subset $P \subseteq X$ using an instance selection matrix IX (a diagonal matrix, diag(IX) = d(X)), where d(X) is a indicator vector, if xi ∈ P, d(X)i = 1, otherwise 0. Accordingly, we define J (P) as an instance evaluation function to measure the P as follows:

$$\mathcal{P}_* = \arg\max_{p \subseteq x} J(P) \quad s.t\,|P| = m$$

where $|\cdot|$ denotes the cardinality of the instance set, and m is the number of instances to be selected from X (i.e., size of DIP). The objective function in Eq. (1) represents that instances selected for MIL P∗ should have maximum discriminative power in the new mapping space.
Bag Mapping Must-Link. Because each bag Bi is associated with a class label (positive or negative), the selected DIP should ensure that bags B i with the same label are similar to each other in the mapping space. (b) Bag Mapping Cannot-Link. For bags with different class labels in the mapping space, they should represent the disparity between them. Accordingly, DIP evaluation criteria could be measured as

$$J(P) = \frac{1}{2}\sum_{i,j} \left\| I_x B_j^{\Phi} \right\|^2 \quad Q_{i,j}$$

After each Bi ∈ B is mapped to B i based on the optimal DIP, any generic single-instance learner can be applied for multi-instance learning.

V. LIMITATIONS OF EXISTING ALGORITHMS

Although various algorithms for saliency detection have been presented in the past few years, and good performance for predicting human fixations in viewing images has been achieved in some circumstances, there are still limitations. Some others highlight object boundaries but fail to detect the whole target region. There are still others emphasizing smaller salient label other than the whole desired one. This inconsistency makes the detection task less fulfilled and limits the usefulness in certain applications. Each characteristic raises different challenges. When instances are grouped in bags, predictions can be performed at two levels: bags-level or instance-level. These two tasks have different misclassification costs therefore algorithms are often better suited for only one of them. Bag composition, such as the pro- portion of instances from each class and the relation between in- stances, also affects the performance of MIL methods. The source of ambiguity on instance labels is another important factor to consider. This ambiguity can be related to label noise as well as to instances not belonging to clearly defined classes. Finally, the shape of positive and negative distributions affects MIL algorithms depending on their assumptions about the data. The bag mapping methods that rely on instance selection are able to prune the instance space; however, it may be difficult to distinguish between the instances in the new bag mapping space. Therefore, designing efficient selection and instance pruning techniques for discriminative bag mapping is important.

MILES does not define an explicit mechanism for instance prototype selection because the IIP is composed of all the instances in the training bags. Once the IIP is formed, MILES maps each bag into a feature space defined by the IIP using a bag-instance similarity measure. However, MILES might potentially map multi-instance learning into a high-dimensionality problem, because the dimensions of the mapping feature space depend on the size of IIP, i.e., the number of instances in training bags. To address this issue, MILIS selects only one instance from each positive bag to prune the instance space. The instance that is most likely to be positive in each positive bag is selected for the IIP, and this is determined by the likelihood of whether an instance is positive using the distributions of all the instances in negative bags. Once constructed, the IIP is used to map each bag into a new bag-level feature space so that a traditional

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 7, Issue 8, August 2018

classifier can be directly employed for further learning. A further type of IIP construction that consists of all the instances within all positive bags along with the clustering centres of instances in negative bags. IIP instance selection is not directly tied to the underlying MIL learning problem, it is difficult to guarantee that the selected instances will be distinguishable from each other in the new bag mapping space.

The propose system multi-Instance learning, the supervised algorithm trains not from single instances but using a group of instances at a time. This group is usually called bags. The multiple-instance binary classification, a bag may be labelled negative if all the instances in it are negative. On the other hand, a bag is labelled positive if there is at least one instance in it which is positive. Most previous multiple-instance learning (MIL) algorithms are developed based on the assumption that a bag mapping is positive if and only if at least one of its instances is positive. Another problem labelling ambiguity in real-world applications. The propose a new algorithm called Multi Cluster Disambiguation-Based Instance Learning (MCDIL), this method to identify the true positive instances in the positive bags using instance-level clustering. The instance-level clustering and bag-level classification, to convert the multi instance learning problem into a normal Single-Instance Learning (SIL) problem that can be solved by familiar SIL algorithms, such as support vector machine. improve the prediction accuracy of the bag labels. In the context of MIL, MCDIL essentially refers to identifying the true positive instances in the positive bags.

VI. CONCLUSION

The most serious problem encumbering the advance of multi-instance learning is that there is only one popularly used real-world benchmark data, i.e. the elephant, tiger, fox data sets. In this paper many researchers have tried to introduce new issues to the research scope of multi-instance learning, which has been surveyed. keeping an eye on applications may not only help us determine the value of extensions of multi-instance learning, but also help us obtain data sets and stimulating problems for multi-instance learning. Finally, experiments are conducted to compare the techniques and disadvantages of 10 state-of-the-art MIL methods papers on selected problem characteristics.

REFERENCES

[1]. T. G. Dietterich, R. H. Lathrop, and T. Lozano-P´erez, “Solving the Multiple Instance Problem with Axis-Parallel Rectangles,” Artificial Intelligence, vol. 89, no. 1-2, pp. 31–71, 1997.
[2]. R. Hong, W. Meng, G. Yue, D. Tao, X. Li, and X. Wu, “Image annotation by multiple-instance learning with discriminative feature mapping and selection,” IEEE Trans. Cybern., vol. 44, no. 5, pp. 669–680, 2014.
[3]. H. Yuan, M. Fang, and X. Zhu, “Hierarchical sampling for multi-instance ensemble learning,” IEEE Trans. Knowl. Data Eng., vol. 25, no. 12, pp. 2900–2905, 2013.
[4]. Q. Wang, Y. Yuan, P. Yan, and X. Li, “Saliency detection by multiple instance learning,” IEEE Trans. Cybern., vol. 43, no. 2, pp. 660–672, 2013.
[5]. J. Wu, X. Zhu, C. Zhang, and P. Yu, “Bag constrained structure pattern mining for multi-graph classification,” IEEE Trans. Knowl. Data Eng., vol. 26, no. 10, pp. 2382–2396, 2014.
[6]. J. Amores, “Multiple instance classification: Review, taxonomy and comparative study,” Artif. Intell., vol. 201, pp. 81–105, 2013.
[7]. Z. Fu, A. Robles-Kelly, and J. Zhou, “Milis: Multiple instance learning with instance selection,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 33, no. 5, pp. 958–977, 2011.
[8]. J. Foulds and E. Frank, “Revisiting multiple-instance learning via embedded instance selection,” in AI, 2008, pp. 300–310.
[9]. J. Wu, Z. Hong, S. Pan, X. Zhu, C. Zhang, and Z. Cai, “Multi-graph learning with positive and unlabeled bags,” in SDM, 2014, pp. 217–225.
[10]. D. Nguyen, C. Nguyen, R. Hargraves, L. Kurgan, and K. Cios, “mi-ds: Multiple-instance learning algorithm,” IEEE Trans. Cybern., vol. 43, no. 1, pp. 143–154, 2013.
[11]. [11] V. Cheplygina, D. M. Tax, and M. Loog, “Multiple instance learning with bag dissimilarities,” Pattern Recogn., vol. 48, no. 1, pp. 264 – 275, 2015.