Ensembles of Multi-Instance Learners

Zhi-Hua Zhou and Min-Ling Zhang

National Laboratory for Novel Software Technology,
Nanjing University, Nanjing 210093, China
zhouzh@nju.edu.cn
http://cs.nju.edu.cn/people/zhouzh/

Abstract. In multi-instance learning, the training set comprises labeled *bags* that are composed of unlabeled instances, and the task is to predict the labels of unseen bags. Through analyzing two famous multi-instance learning algorithms, this paper shows that many supervised learning algorithms can be adapted to multi-instance learning, as long as their focuses are shifted from the discrimination on the instances to the discrimination on the bags. Moreover, considering that ensemble learning paradigms can effectively enhance supervised learners, this paper proposes to build ensembles of multi-instance learners to solve multi-instance problems. Experiments on a real-world benchmark test show that ensemble learning paradigms can significantly enhance multi-instance learners, and the result achieved by EM-DD ensemble exceeds the best result on the benchmark test reported in literature.

1 Introduction

The term *multi-instance learning* was coined by Dietterich et al. [11] when they were investigating the problem of drug activity prediction. In this learning framework, the training set is composed of many *bags* each contains many instances. A bag is positively labeled if it contains at least one positive instance. Otherwise it is negatively labeled. The task is to learn some concept from the training bags for correctly labeling unseen bags. This task is very difficult because unlike supervised learning where all the training instances are labeled, here the labels of the individual instances are unknown. It has been shown that learning algorithms ignoring the characteristics of multi-instance learning could not work well in this scenario [11].

The PAC-learnability of multi-instance learning has been studied by many researchers [2][3][5][13], and some important results, such as 'if the instances in the bags are not independent then APR learning under the multi-instance learning framework is NP-hard', have been obtained. The most famous multi-instance learning algorithm is Diverse Density [14] which has been applied to several applications including stock prediction [14], natural scene classification [15], and content-based image retrieval [20]. There are also many other practical algorithms, such as Citation-kNN [18], Relic [17], ID3-MI [8], RIPPER-MI [8], EM-DD [21], BP-MIP [23], etc. Recently, multi-instance regression with real-valued outputs begins to be studied [1][16]. It is worth noting that multi-instance

learning has also attracted the attention of the ILP community. It has been suggested that multi-instance problems could be regarded as a bias on inductive logic programming, and the multi-instance paradigm could be the key between the propositional and relational representations, being more expressive than the former, and much easier to learn than the latter [9].

In this paper, two famous multi-instance learning algorithms, i.e. Diverse Density and Citation-kNN, are analyzed, which suggests that many supervised learning algorithms can be adapted to multi-instance learning as long as they attempt to discriminate the bags instead of the instances. Then, considering that ensemble learning paradigms that train multiple learners to solve a problem can effectively improve the generalization ability in supervised learning [10], this paper proposes to build multi-instance ensembles to solve multi-instance problems. Experiments on a real-world benchmark data set show that current multi-instance learners can be significantly enhanced by ensemble learning paradigms. Moreover, it is observed that the ensemble of a specific multi-instance learner, i.e. EM-DD, exhibits the best performance up to date on the benchmark test.

The rest of this paper is organized as follows. Section 2 analyzes the Diverse Density algorithm and the Citation$k$NN algorithm. Section 3 proposes to build multi-instance ensembles. Section 4 presents the experimental results. Finally, Section 5 summarizes the contributions of this paper.

2 Adapt Supervised Algorithms to Multi-Instance Learning

When proposing the notion of multi-instance learning, Dietterich et al. [11] raised an open problem, i.e. designing multiple instance modifications for popular machine learning algorithms. In fact, multi-instance versions of many machine learning algorithms have been developed in recent years [8][17][18][23]. However, there is no general rule indicating how to do such a modification.

Usually, the focus of a supervised learning algorithm is to discriminate the instances, which is feasible since all training instances are labeled in supervised scenario. But in multi-instance learning, it is infeasible to build a model through discriminating training instances because none of them is labeled. Moreover, if the label of a bag is simply regarded as the label of its instances, i.e. to believe that positive bag contains only positive instances and negative bag contains only negative instances, then the learning task may be very difficult although every training instance holds a label now. This is because the positive noise may be extremely high [1], as indicated by [11]. Therefore, whether it is possible to discriminate the training instances or not is the principal difference between supervised and multi-instance learning.

In this section we claim that many supervised learning algorithms can be adapted to multi-instance learning, as long as they shift the focus from the

[1] Consider that a positive bag may contain hundreds or even thousands of negative instances but only one positive instance.

discrimination on the instances to the discrimination on the bags. We illustrate that two multi-instance learning algorithms, i.e. Diverse Density and Citation-kNN, can be derived from standard Bayesian classifier and k-nearest neighbor algorithm according to our claim. These two algorithms are chosen to analyze because Diverse Density is the most famous multi-instance learning algorithm at present, and Citation-kNN had achieved the best result on the real-world multi-instance benchmark test [18] before EM-DD, a variant of Diverse Density, was proposed.

2.1 Diverse Density

The Diverse Density algorithm [14] regards each bag as a manifold, which is composed of many instances, i.e. feature vectors. If a new bag is positive then it is believed to intersect all positive feature-manifolds without intersecting any negative feature-manifolds. Intuitively, *diverse density* at a point in the feature space is defined to be a measure of how many different positive bags have instances near that point, and how far the negative instances are from that point. Thus, the task of multi-instance learning is transformed to search for a point in the feature space with the *maximum diverse density*.

It is evident that the key of the Diverse Density algorithm lies in the formal definition of the *maximum diverse density*, which is the objective to be optimized by the algorithm. Below we show that such a definition can be achieved through adapting standard Bayesian classifier according to the rule, i.e. shifting the focus from discriminating the instances to discriminating the bags.

Given data set D and a set of class labels, i.e. $C = \{c_1, c_2, \cdots, c_t\}$, to be predicted, the posterior probability of the class can be estimated according to the Bayes rule as shown in Eq. 1.

$$\Pr(C\,|D\,) = \frac{\Pr(D\,|C\,)\Pr(C)}{\Pr(D)} \tag{1}$$

What we want is the class label with the maximum posterior probability, as indicated in Eq. 2, where *Obj* denotes the objective.

$$\mathrm{Obj} = \arg\max_{1\le k\le t} \Pr(c_k\,|D\,) = \arg\max_{1\le k\le t} \frac{\Pr(D\,|c_k\,)\Pr(c_k)}{\Pr(D)} \tag{2}$$

Considering that $\Pr(D)$ is a constant which can be dropped, and $\Pr(c_k)$ can also be dropped if we assume uniform prior, then Eq. 2 can be simplified as Eq. 3.

$$\mathrm{Obj} = \arg\max_{1\le k\le t} \Pr(D\,|c_k\,) \tag{3}$$

Eq. 3 is enough when the goal is to discriminate the instances. But for discriminating the bags, it is helpful to consider $D = \{B_i^+, \cdots, B_m^+, B_1^-, \cdots, B_n^-\}$ where B_i^+ denotes the i-th positive bag while B_j^- denotes the j-th negative bag.

Then, Eq. 3 can be re-written as Eq. 4 assuming that the bags are conditionally independent.

$$\begin{aligned}\text{Obj} &= \arg\max_{1\le k\le t} \Pr\left(\{B_i^+, \cdots, B_m^+, B_1^-, \cdots, B_n^-\} \,|c_k\right) \\ &= \arg\max_{1\le k\le t} \prod_{1\le i\le m} \Pr\left(B_i^+ \,|c_k\right) \prod_{1\le j\le n} \Pr\left(B_j^- \,|c_k\right) \end{aligned} \tag{4}$$

Now apply the Bayes rule to the right-hand of Eq. 4, we get Eq. 5.

$$\text{Obj} = \arg\max_{1\le k\le t} \prod_{1\le i\le m} \frac{\Pr\left(c_k \left| B_i^+\right.\right) \Pr\left(B_i^+\right)}{\Pr\left(c_k\right)} \prod_{1\le j\le n} \frac{\Pr\left(c_k \left| B_j^-\right.\right) \Pr\left(B_j^-\right)}{\Pr\left(c_k\right)} \tag{5}$$

Considering that $\prod_{1\le i\le m} \Pr\left(B_i^+\right) \prod_{1\le j\le n} \Pr\left(B_j^-\right)$ is a constant which can be dropped, and reminding that $\Pr\left(c_k\right)$ can be dropped as that has been done in Eq. 3 because we assume uniform prior, then Eq. 5 can be simplified as Eq. 6.

$$\text{Obj} = \arg\max_{1\le k\le t} \prod_{1\le i\le m} \Pr\left(c_k \left| B_i^+\right.\right) \prod_{1\le j\le n} \Pr\left(c_k \left| B_j^-\right.\right) \tag{6}$$

It is interesting that Eq. 6 is neither more nor less than the formal definition of the *maximum diverse density* which is optimized by the Diverse Density algorithm [14]!

2.2 Citation-*k*NN

The Citation-*k*NN algorithm [18] is a nearest neighbor style algorithm, which borrows the notion of citation of scientific references in the way that a bag is labeled through analyzing not only its neighboring bags but also the bags that regard the concerned bag as a neighbor.

Nevertheless, it is evident that for any nearest neighbor style algorithm, the key lies in the definition of the distance metric which is utilized to measure the distance between different objects. Below we show that the key of Citation-*k*NN, i.e. the definition of the *minimal Hausdorff distance*, can be achieved through adapting standard *k*-nearest neighbor algorithm according to the rule, i.e. shifting the focus from discriminating the instances to discriminating the bags.

In standard *k*-nearest neighbor algorithm, each object, or instance, is regarded as a feature vector in the feature space. For two different feature vectors, i.e. a and b, the distance between them can be written as Eq. 7. Usually $\|a-b\|$ is realized as the Euclidean distance.

$$\text{Dist}\left(a, b\right) = \|a - b\| \tag{7}$$

When the goal is to discriminate the instances, Eq. 7 is enough to be instantiated. But if the goal is to discriminate the bags, then Eq. 7 must be extended because now we should measure the distance between different bags.

Suppose we have two different bags, i.e. $A = \{a_1, a_2, \cdots, a_m\}$ and $B = \{b_1, b_2, \cdots, b_n\}$ where a_i $(1 \leq i \leq m)$ and b_j $(1 \leq j \leq n)$ are the instances. It is obvious that they can be regarded as two feature vector sets, where each a_i $(1 \leq i \leq m)$ or b_j $(1 \leq j \leq n)$ is a feature vector in the feature space. Therefore, the problem of measuring the distance between different bags is in fact the problem of measuring the distance between different feature vector sets.

Geometrically, a feature vector set can be viewed as a group of points enclosed in a contour in the feature space. Thus, an intuitive way to measure the distance between two feature vector sets is to define their distance as the distance between their nearest feature vectors, as illustrated in Fig. 1.

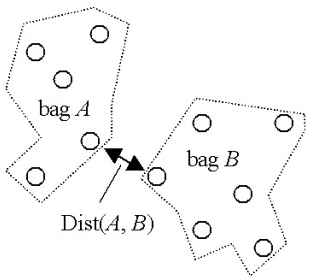

Fig. 1. An intuitive way to define the distance between bags

Formally, such a distance metric can be written as Eq. 8.

$$\mathrm{Dist}\,(A, B) = \min_{\substack{1 \leq i \leq m \\ 1 \leq j \leq n}} (\mathrm{Dist}\,(a_i, b_j)) = \min_{a \in A} \min_{b \in B} \|a - b\| \tag{8}$$

It is interesting that the right-hand of Eq. 8 is neither more nor less than the formal definition of the *minimum Hausdorff distance*, which is employed by the Citation-*k*NN algorithm [18] to measure the distance between different bags!

3 Multi-Instance Ensemble

Ensemble learning paradigms train multiple versions of a base learner to solve a problem. Since ensembles are usually more accurate than single learners, one of the most active areas of research in supervised learning has been to study paradigms for constructing good ensembles [10].

Since we have shown in Section 2 that many supervised learning algorithms can be adapted to multi-instance learning, a consequent exciting idea is to see whether ensemble learning paradigms can be used to enhance multi-instance learners. Here we call ensemble of multi-instance learners as multi-instance ensemble.

During the past years, diverse ensemble learning algorithms have been developed, such as Bagging [6], Arc-x4 [7], AdaBoost [12], MultiBoost [19], GASEN [22], etc. In this section, we use a relatively simple algorithm, i.e. Bagging, to build the multi-instance ensembles.

Bagging employs bootstrap sampling to generate several training sets from the original training set and then trains component learners, i.e. multiple versions of the base learner, from each generated training set. The predictions of the component learners are combined via majority voting. The Bagging algorithm is shown in Table 1, where T bootstrap samples $S_1, S_2, \cdots, S_T$ are generated from the training set S and an component learner L_t is trained from each S_t, an ensemble L^* is built from $L_1, L_2, \cdots, L_T$ whose output is the class label receiving the most number of votes, x is the input feature vector, and Y is the set of class labels.

Table 1. The Bagging algorithm

Input: training set S, base learner L, trials of bootstrap sampling T
Output: ensemble L^*
Process:
for $t = 1$ to T {
S_t = bootstrap sample from S
$L_t = L(S_t)$
}
$$L^*(x) = \arg\max_{y \in Y} \sum_{t:\, L_t(x)=y} 1$$

We attempt to build multi-instance ensembles for four different base learners, i.e. Iterated-discrim APR [11], Diverse Density [14], Citation-kNN [18], and EM-DD [21]. The reason for choosing Diverse Density and Citation-kNN have been introduced in Section 3. Here we briefly explain why the other two algorithms are chosen.

Iterated-discrim APR is the best Axis-Parallel Rectangle (abbreviated as APR) algorithm proposed by Dietterich et al. [11], which attempts to search for appropriate axis-parallel rectangles constructed by the conjunction of the features. Dietterich et al. [11] indicated that since the APR algorithms had been optimized to the *Musk* data, i.e. the only real-world multi-instance benchmark data until now, the performance of Iterated-discrim APR might be the upper bound of this benchmark test.

EM-DD [21] is a recent development in multi-instance learning, which combines the EM and Diverse Density algorithms. It converts the multi-instance problem to a single-instance setting by using EM to estimate the instance which is responsible for the label of the bag. The best performance on the real-world multi-instance benchmark test until now, i.e. predictive error rate as small as 3.2% on *Musk1* and 4.0% on *Musk2*, are achieved by this algorithm [21]. Note

that the performance of EM-DD has already exceeded the upper bound of this benchmark test anticipated by Dietterich et al. [11].

4 Experiments

The experiments are performed on the *Musk* data, which is the only real-world benchmark test data for multi-instance learners at present.

The *Musk* data were generated in Dietterich et al.'s research on drug activity prediction [11]. Here each molecule is regarded as a bag, and its alternative low-energy shapes are regarded as the instances in the bag. A positive bag corresponds to a molecule qualified to make a certain drug, that is, at least one of its low-energy shapes could tightly bind to the target area of some larger protein molecules such as enzymes and cell-surface receptors. A negative bag corresponds to a molecule not qualified to make a certain drug, that is, none of its low-energy shapes could tightly bind to the target area.

In order to represent the shapes, a molecule is placed in a standard position and orientation and then a set of 162 rays emanating from the origin is constructed so that the molecular surface is sampled approximately uniformly. There are also four features that represented the position of an oxygen atom on the molecular surface. Therefore each instance in the bags is represented by 166 continuous attributes.

There are two data sets, i.e. *Musk1* and *Musk2*, both of which are publicly available from the UCI Machine Learning Repository [4]. *Musk1* contains 47 positive bags and 45 negative bags, and the number of instances contained in each bag ranges from 2 to 40. *Musk2* contains 39 positive bags and 63 negative bags, and the number of instances contained in each bag ranges from 1 to 1,044. Detailed information on the *Musk* data is tabulated in Table 2.

Table 2. The *Musk* data (72 molecules are shared in both data sets)

Data set	Dim.	Bags			Instances	Instances per bag		
		Total	Musk	Non-musk		Min	Max	Ave.
Musk1	166	92	47	45	476	2	40	5.17
Musk2	166	102	39	63	6,598	1	1,044	64.69

Ten-fold cross validation is performed on each *Musk* data set. In each fold, Bagging is employed to build an ensemble for each of the four base multi-instance learners, i.e. i.e. Iterated-discrim APR, Diverse Density, Citation-kNN, and EM-DD. Each ensemble comprises five versions of the base learner. The predictive error rates of the ensembles are shown in Table 3. For comparison, the best results of the single multi-instance learners reported in the literatures [11][14][18][21] are also included in Table 3.

Table 3. Predictive error rates (%) of single or ensembled multi-instance learners

Algorithm	*Musk1*		*Musk2*	
	Single	Ensemble	Single	Ensemble
Iterated-discrim APR	7.6	7.2	10.8	6.9
Diverse Density	11.1	8.2	17.5	11.0
Citation-kNN	7.6	5.2	13.7	12.9
EM-DD	3.2	3.1	4.0	3.0

Table 3 shows that Bagging can significantly improve the generalization ability of all the investigated multi-instance learners.[2] It is impressive that even the strongest multi-instance learner, i.e. EM-DD, can be enhanced by such a relatively simple ensemble learning algorithm. In fact, the EM-DD ensemble achieves the best performance up to date on both the *Musk* data sets, i.e. predictive error rate 3.1% on *Musk1* and 3.0% on *Musk2*.

Since the process of building ensemble of multi-instance learners has nothing being geared to any specific data, we believe that such a paradigm can be applied to any multi-instance problems. It is also reasonable to anticipate that such a paradigm may return more profit on difficult problems where no single multi-instance learners works very well. Moreover, the experiments reported in this section also suggest ensemble learning paradigms be investigated in more scenarios, not to be limited in supervised learning.

5 Conclusion

When formalizing the notion of multi-instance learning, Dietterich et al. [11] raised an open problem, i.e. designing multiple instance modifications for popular machine learning algorithms. Although multi-instance versions of many machine learning algorithms have been developed in recent years, there is no general rule indicating how to do such a modification until now.

This paper claims that many supervised learning algorithms can be adapted to multi-instance learning through shifting their focuses from the discrimination on instances to the discrimination on bags. In order to illustrate the feasibility of such kind of adaptation, two famous multi-instance algorithms, i.e. Diverse Density and Citation-kNN, are analyzed in this paper, which shows that they can be derived from standard Bayesian classifier and k-nearest neighbor algorithm, respectively, through shifting the focus.

Designing multi-instance learning algorithms with strong generalization ability is always an important issue in this area. Considering that many supervised learning algorithms can be adapted to multi-instance learning, and ensemble

[2] The results of the single multi-instance learners in Table 3 are the best results reported by their authors [11][14][18][21]. In our implementation, the performance of the single learners are slightly worse than these best results.

learning paradigms can effectively enhance supervised learners, this paper claims to build multi-instance ensembles to solve multi-instance problems.

Experiments shows that all the investigated multi-instance learners can be enhanced by a relatively simple ensemble learning algorithm, and the best result up to date on the real-world benchmark test of multi-instance learners is achieved by EM-DD ensemble. The experiments not only support our claim that building multi-instance ensembles is a good choice for solving multi-instance problems, but also suggest ensemble learning paradigms be investigated in more scenarios, not to be limited in supervised learning.

Acknowledgement

This work was supported by the National Natural Science Foundation of China under the Grant No. 60105004, and the Natural Science Foundation of Jiangsu Province under the Grant No. BK2001406.

References

1. Amar, R.A., Dooly, D.R., Goldman, S.A., Zhang, Q.: Multiple-instance learning of real-valued data. In: Proceedings of the 18th International Conference on Machine Learning, Williamstown, MA (2001) 3–10
2. Auer, P.: On learning from multi-instance examples: empirical evaluation of a theoretical approach. In: Proceedings of the 14th International Conference on Machine Learning, Nashville, TN (1997) 21–29
3. Auer, P., Long, P.M., Srinivasan, A.: Approximating hyper-rectangles: learning and pseudo-random sets. Journal of Computer and System Sciences **57** (1998) 376–388
4. Blake, C., Keogh, E., Merz, C.J.: UCI repository of machine learning databases [http://www.ics.uci.edu/∼mlearn/MLRepository.html], Department of Information and Computer Science, University of California, Irvine, CA (1998)
5. Blum, A., Kalai, A.: A note on learning from multiple-instance examples. Machine Learning **30** (1998) 23–29
6. Breiman, L.: Bagging predictors. Machine Learning **24** (1996) 123–140
7. Breiman, L.: Bias, variance, and arcing classifiers. Technical Report 460, Statistics Department, University of California, Berkeley, CA (1996)
8. Chevaleyre, Y., Zucker, J.-D.: Solving multiple-instance and multiple-part learning problems with decision trees and rule sets. Application to the mutagenesis problem. In: Stroulia, E., Matwin, S. (eds.): Lecture Notes in Artificial Intelligence, Vol. 2056. Springer, Berlin (2001) 204–214
9. De Raedt, L.: Attribute-value learning versus inductive logic programming: the missing links. In: Page, D. (ed.): Lecture Notes in Artificial Intelligence, Vol. 1446. Springer, Berlin (1998) 1–8
10. Dietterich, T.G.: Machine learning research: four current directions. AI Magazine **18** (1997) 97–136
11. Dietterich, T.G., Lathrop, R.H., Lozano-Pérez, T.: Solving the multiple-instance problem with axis-parallel rectangles. Artificial Intelligence **89** (1997) 31–71
12. Freund, Y., Schapire, R.E.: A decision-theoretic generalization of on-line learning and an application to boosting. In: Proceedings of the 2nd European Conference on Computational Learning Theory, Barcelona, Spain (1995) 23–37

13. Long, P.M., Tan, L.: PAC learning axis-aligned rectangles with respect to product distributions from multiple-instance examples. Machine Learning **30** (1998) 7–21
14. Maron, O., Lozano-Pŕez, T.: A framework for multiple-instance learning. In: Jordan, M.I., Kearns, M.J., Solla, S.A. (eds.): Advances in Neural Information Processing Systems, Vol. 10. MIT Press, Cambridge, MA (1998) 570–576
15. Maron, O., Rantan, A.L.: Multiple-instance learning for natural scene classification. In: Proceedings of the 15th International Conference on Machine Learning, Williamstown, MA (2001) 341–349
16. Ray, S., Page, D.: Multiple instance regression. In: Proceedings of the 18th International Conference on Machine Learning, Williamstown, MA (2001) 425–432
17. Ruffo, G.: Learning single and multiple instance decision trees for computer security applications. PhD dissertation, Department of Computer Science, University of Turin, Torino, Italy (2000)
18. Wang, J., Zucker, J.-D.: Solving the multiple-instance problem: a lazy learning approach. In: Proceedings of the 17th International Conference on Machine Learning, San Francisco, CA (2000) 1119–1125
19. Webb, G.I.: MultiBoosting: a technique for combining Boosting and Wagging. Machine Learning **40** (2000) 159–196
20. Yang, C., Lozano-Pérez, T.: Image database retrieval with multiple-instance learning techniques. In: Proceedings of the 16th International Conference on Data Engineering, San Diego, CA (2000) 233–243
21. Zhang, Q., Goldman, S.A.: EM-DD: an improved multi-instance learning technique. In: Dietterich, T.G., Becker, S., Ghahramani, Z. (eds.): Advances in Neural Information Processing Systems, Vol. 14. MIT Press, Cambridge, MA (2002) 1073–1080
22. Zhou, Z.-H., Wu, J., Tang, W.: Ensembling neural networks: many could be better than all. Artificial Intelligence **137** (2002) 239–263
23. Zhou, Z.-H., Zhang, M.-L.: Neural networks for multi-instance learning. In: Proceedings of the International Conference on Intelligent Information Technology, Beijing, China (2002) 455–459