Handwritten Digits Recognition Using Multiple Instance Learning

Yuan hanning
International School of Software
Beijing Institute of technology
Beijing, China
Email: yhn6@bit.edu.cn

Wang Peng
Department of Systems Innovations
School of Engineering, the Univ. of Tokyo
Tokyo, Japan
Email: pwang@save.sys.t.u-tokyo.ac.jp

Abstract **Now more and more heterogeneous handwritten digits data sets appear into sight. But traditional handwritten digits recognition algorithms are usually based on the homomorphism data sets. For solving the problem that handwritten digits data sets of different feature spaces can't compute, we constructed heterogeneous handwritten digits representation model based on multiple instance learning (MIL) where a bag contains handwritten digits data from different feature spaces. Handwritten digits classification algorithms (HB and HeterMIL) are designed and compared for handwritten digits recognition. Experiment results confirmed that the heterogeneous handwritten digits data representation model and recognition algorithms can solve the heterogeneous handwritten digits recognition effectively.**

Keywords Multipe instance learning; heterogeneous handwritten digits; classification

I. INTRODUCTION

Multi-instance learning (MIL), originally proposed by Dietterichet.al[1] for drug activity predictions, represents a special type of machine learning task where a label is only available for a bag, which contains a number of instances, but no label is available for individual instances inside each bag. In another word, the label of each individual instance in a bag is unobserved. Each instance within a bag is described by a vector of features. Assume the given of a number of labeled bags, the goal of the MIL is to construct a learner to predict whether a previously unseen bag to be positive or negative.

In recent years, many MI algorithms have been proposed in the literature[2].It has recently attracted a lot of attentions for applications such as drug activity prediction [1], content based image retrieval [3], web index page recommendation [4], stock selection, and natural scene classification [5].

However, in most cases, the proposed algorithms were often under the assumption that all instances are in the same feature space [6], which we named this kind of MI as Homogeneous Multi-instance Learning (HomoMIL) in this thesis. However, in many real world MI tasks, instances often come from different feature spaces, which we named this kind of MI as Heterogeneous Multi-instance Learning (HeterMIL) in this thesis. Common HomoMIL algorithms can't solve HeterMIL for almost all MI algorithms learned concept form examples by computing or comparing instances which are in the same feature space while in HeterMIL the computing and comparing instances come from different feature space.

The aim of this thesis is to present a study of HeterMIL algorithms for handwritten digits recognition. Heterogeneous handwritten digits representation model based on multiple instance learning is constructed and HeterMIL classification algorithms are designed. Experiments on a collection of real-world datasets are performed in an attempt to address some specific questions regarding HeterMIL problems.

II. HETEROGENEOUS REPRESENTATION MODEL AND CLASSIFICATION FOR HANDWRITTEN DIGITS

A. heterogeneous MI representation model

Different to MIL, instances of a bag in HeterMIL probably come from different feature spaces (as show in figure 1). Each feature space is described by a vector of features. We called each features vector as a modality. Similar to MIL, in HeterMIL a bag is positively labeled if it contains at least one positive instance; otherwise it is labeled as a negative bag. The task is to learn some concept from the training set.

Use B_i denotes the ith bag. In each bag, the number of instance modality is one or more than one, the set of modalities is denoted as M. the number of bags is |B| ,the number of positive big bags is| B^+ | ,the number of negative bags is | B^- |. HeterMIL can be described as Figure 1.

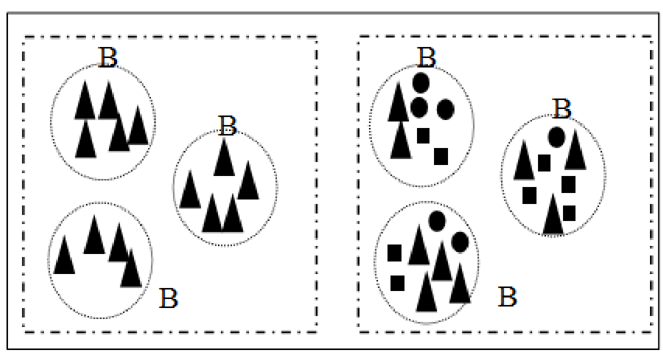

Figure 1. Left picture shows the homogeneous MIL, right picture shows an H-MIL dataset with 3 different kinds of modalities. Triangle denotes an instance of modality 1. Rectangle denotes an instance of modality 2 Round denotes an instance of modality 3

B. classfication Algorithms for Heterogeneous MIL

One solution for the problem is to learn concept respectively in each modality of instances by dividing the instances into several groups according to their modality. The final decision is the combination of all decision on each modality instances. The key of the solution is to obtain the instance label information in training datasets. Similar to MIL, HeterMIL follows the same assumption that a bag is positively labeled if it contains at least one positive instance; otherwise it is labeled as a negative bag. So the label information of instances in negative bags is clear that all instances in negative bags are negative. But the label information of instances in positive bags is ambiguous. In this thesis, we have two strategies to obtain the instance label information of positive bags. One is directly propagate the bag label to instances in it and methods using this strategy are named as heterogeneous baseline methods (HB). The other strategy is heuristically find the most probably positive instance in each positive bag and methods using this strategy are named as as heterogeneous heuristic MIL methods (HHMI).

The core idea of both methods is to divide bag into several small bags according to instance modality and to propagate the bag label to small bags of it. However, according to MI standard assumption, directly propagating the bag label to instances of it would lead to incorrectly label instances in positive bags which will limit the performance of learning.

So the main technical challenge in HeterMIL is caused by bag labeling constraints (i.e. a positive bag contain at least one positive sample)[7]. For the labels of instances in positive bags are ambiguous, the most difficult work is to find the true positive instances. To find true positive instances of positive bags, a heterogeneous heuristic multiple instance learning method (HHMIL) is designed.

Different to baseline methods which directly prorogate the bag label to instance, HHMIL tried to find true positive instances in positive bags heuristically. Firstly, bag labels are directly prorogated to instances and instances are divided into some instances groups according to instance modality. Secondly SVM is trained respectively in each instances group modality and trained SVM is used to compute the probability of all instances in positive bags to be positive in each instance group modality. If the probability of an instance of positive bag to be positive is lower than 50%, the instance is labeled as negative. Otherwise the instance is labeled as negative. Considering the bag constraint of MIL, if all the instances in a positive bag are all labeled as negative, the instance with the biggest probability is labeled as positive instance. The step repeated literately until no instance label changes.

When classifying a bag, for each instance in the bag, trained SVM is chosen according to the modality of instance to classify the instance and the final decision is the combine of all instance classification result according to decision rules. Followed is the algorithm of HHMIL.

Algorithms: A heterogeneous heuristic multiple instance learning method (HHMIL)

Input: training dataset T^M the number of modality category is M, base MI learning Algorithm Ln

Output: The final learner L

Process

Iteration of instances selection can be expressed in following five steps:

The first time of iteration:

1. Build learner L_1 of T^M

2. Select characteristic instances: Classify the T^M by L_1, if the prediction of an instance is right then it put into subT_1^M

Iterations from the second time:

3. Initialize T_n^M by subT_n^M (n=1,2,3…)

4. Build a new classifier L_{n+1} of T_n^M

5. Classify the T^M by L_{n+1}, if the prediction of an instance is right then it put into subT_{n+1}^M

6. Repeat step 3 to step 5 until approach condition of convergence.

7. Output the final Learner L_{n+1}

Here, Ln defines the SVM classifier of the n^{th} times iteration. And T_n^M represents the data set of the n^{th} times iteration of M modality. Define the selection data set of M modality as subT_n^M. Except the first time, the T_{n+1}^M will be initialized as $subT_n^M$. M modality is taken as an example of instances selection.

III. EXPERIMENT AND RESULT ANALYSIS

A. Data source

We use a multiple features dataset of handwritten digits from the UCI Machine Learning Repository to generate dataset for heterogeneous MIL.

A random sampling method was designed to generate the specific dataset for Heterogeneous MIL. There are two integer parameters randomly generated, which are dataset number and line number. The instances in a bag are sampled from lines of the feature set. Every feature set is represented as a modality. The modality of each instance is tabbed as well when it is sampled. The types of modality in a generated train set determined the name of it. For example, if a train set contains 2 varieties of modality, the set name will contain the key words “2M”.And in our experiment, the digits is only considering from 0 to 4 for data balance.

In order to make experimental result credibility and referable for the further research, the train sets we generated is consisted by 2M, 3M, 4M, 5M and 6M.The generated experiment data is shown as Table I.

TABLE I. THE HETEROGENEOUS HANDWRITTEN DIGITS DATA SETS

Data set name	The number of positive instances	Average number of instance in positive bag	The number of negative instances	Average number of instance in negative bag	The Number of all instances	Average number of instance in all bags
HandWrite-2M-1	514	10.28	284	5.68	798	7.98
HandWrite-3M-1	613	12.26	480	9.6	1093	10.93
HandWrite-4M-1	743	14.86	651	13.02	1394	13.94
HandWrite-5M-1	863	17.26	833	16.66	1696	16.96
HandWrite-6M-1	921	18.42	1103	22.06	2024	20.24
HandWrite-2M-2	434	8.68	365	7.3	799	7.99
HandWrite-3M-2	519	10.38	580	11.6	1099	10.99
HandWrite-4M-2	740	14.8	657	13.14	1397	13.97
HandWrite-5M-2	922	18.44	777	15.54	1699	16.99
HandWrite-6M-2	1002	20.04	996	19.92	1998	19.98

B. *Experiment setting*

We implement the methods using Java platform and WEKA machine learning tool[9].10-fold cross validations is used to evaluate the accuracy of Heterogeneous MI learning methods and the times of cross validation is set as 5. In each fold, the training and test data sets are kept the same for all methods. In our experiment, the feature weight values are calculated by using the information gain[8].

Three kinds of decision rules are discussed in the thesis, which are shown as following:

DS_1 rule. This kind of decision rule strictly follows the MI standard assumption. In a test bag, if there is at least one instance of the bag whose probability to be positive is more than 0.5, the bag will labeled as positive otherwise it is labeled as negative.

DS_2 rule. MI collection assumption is used to guide the final decision of bag in this rule. The decision rule assumes that all instances in a bag contribute equally. The final decision of bag label should be made based on the average of the class probabilities of all individual instances.

DS_3 rule. This is a standard assumption based method as well. According to the generation method of train set, the probability of a sampled instance from a specific modality labeled positive (the instance that represents '0') is 1/5. So if more than 20% of the instances in a test bag are predicted to be positive, the bag will labeled as positive otherwise it is labeled as negative.

C. *Experiment and result analysis*

For comparison purposes, we implement three baseline methods, including HB, HB /2 and HB 1/5, and three HHMI methods denoted by HHMI, HHMI1/2, HHMI1/5. HB 1/2 and HB 1/5 are almost are identical to baseline except their decision rule of bag label. The decision rule baseline adopted is DS_1 while HB 1/2 uses DS_2 and HB 1/5 uses DS_3. HHMI 1/2 and HHMI 1/5 are almost are identical to HHMI except their decision rule of bag label. The decision rule baseline adopted is DS_1 while HHMI1/2 1/2 uses DS_2 and HHMI1/5 uses DS_3.

The experimental results of 10 generated data sets computed by HB and HHMI are shown as following in the Table II.

TABLE II. RESULTS OF HB AND HHMI

dataset	HB	HHMI	HB 1/2	HHMI 1/2	HB 1/5	HHMI 1/5
HandWrite -2M-1	73.6	79	70.8	74.8	73.8	**79.8**
HandWrite -3M-1	80	82.4	89.2	**92.4**	87	90.6
HandWrite -4M-1	70.4	76	68.4	81.6	75.4	**82.4**
HandWrite -5M-1	64.2	70.8	83.8	**93.6**	69.8	78.4
HandWrite -6M-1	50.8	53.4	74.8	**86.4**	57.6	61.2
HandWrite -2M-2	86.8	**90.8**	81.6	84.2	87.2	91.4
HandWrite -3M-2	72	79.6	82.2	**90.2**	76.8	85.2
HandWrite -4M-2	57.4	58.8	79.8	**83.6**	60.4	64.8
HandWrite -5M-2	52	59	75.4	**88.8**	59.2	70.4
HandWrite -6M-2	53.6	57	76.6	**90.2**	58.8	65.8

Table III shows the average accuracy of 2 groups of experiments. Table IV is the deeply reduction of Table II. Every value in the Table IV is the average of HB and HHMI. The table header in TABLE II is merged into DS1, DS 2 and DS 3 in Table IV, which corresponding to the different decision rules of bag label.

TABLE III. THE AVERAGE ACCURACY

dataset	baseline	heuristic	baseline1/2	heuristic1/2	baseline1/5	heuristic1/5
2M	80.2	84.9	76.2	79.5	80.5	85.6
3M	76	81	85.7	91.3	81.9	87.9
4M	63.9	67.4	74.1	82.6	67.9	73.6
5M	58.1	64.9	79.6	91.2	64.5	74.4
6M	52.2	55.2	75.7	88.3	58.2	63.5

In Table II, we report experimental comparisons across different Heterogeneous learning methods and different benchmark datasets. For each row, the method with the highest mean accuracy is bold faced.

When comparing HB and HHMIL methods across different benchmark datasets, it is clear that HHMIL does show to be effective to improve the classification accuracy than HB. From Table II, when DS1 is adopted, comparing baseline and heuristic method on 10 dataset, heuristic win baseline 10 times. When DS2 is adopted, HHMI1/2 wins HB1/2 10 times. When DS3 is adopted, HHMI1/5 wins HB1/5 10 times. So it is obvious that no matter which DS rule is chosen the HHMI got a higher accuracy than the HB.

The reason is that HHMI take the bag constraint into account which is ignored in HB.

TABLE IV. THE DEEPLY REDUCTION

dataset	DS_1	DS_2	DS_3
HandWrite -2M-1	5.4	4	6
HandWrite -2M-2	4	2.6	4.2
HandWrite -3M-1	2.4	3.2	3.6
HandWrite -3M-2	7.6	8	8.4
HandWrite -4M-1	5.6	13.2	7
HandWrite -4M-2	1.4	3.8	4.4
HandWrite -5M-1	6.6	9.8	8.6
HandWrite -5M-2	7	13.4	11.2
HandWrite -6M-1	2.6	11.6	3.6
HandWrite -6M-2	3.4	13.6	7
average	4.6	8.32	6.4

Table IV is generated from Table II, which shows the improvement of HHMIL comparing with HB under same decision rule on every data set. The last row in this table is the average improvement of HHMIL comparing with HB under same decision rule on every data set.

Comparing with baseline method, HHMI gets the average increase with 8.32% when it adopts DS_2. The most notable increase even reaches 13.6%. There are 4 accuracies improvement in this column above 10%. Under DS_3, the average improvement reduces about 2%, as 6.4%. There appears only an increase rate above 10%.Under DS_1, major values are at a lower level, the most sharp improvement rate of which is only 7.6%. The table explains HHMI improve accuracy on different levels under different decision rules. It seems that decision rules influence the performance gain in some degree

IV. CONCLUSION

In this thesis, heterogeneous handwritten digits representation model based on multiple instance learning is constructed and two new kinds of heterogeneous MIL methods named HB and HeterMI are put forward. Experimental comparisons across MI learning methods and benchmark data sets demonstrated that HeterMIL achieves better performance than HB for it takes the bag constraints into consideration.

The influence of three different decision rules is compared in this thesis. DS_1 method that based on standard assumption seems not fit for the data set in this experiment. Though DS_2 method performs well, especially in high modality data set which is unexpected, we will not choose DS_2 in future research because the improvement may be caused by the more miss prediction of classifier. However, DS_3 method is very close to nature probability of the original data sets and it is also the most potential method in the future research.

ACKNOWLEDGMENT

The work is supported by National Science Foundation of China(NSFC 71201120) and the Fundamental Research Funds for the Central Universities (2012-IV-053).

REFERENCES

[1]Dietterich T G, Lathrop R H, Lozano-Pérez T. Solving the multiple-instance problem with axis parallel rectangles [J]. Artificial Intelligence , 1997, 89(1-2): 31-71.

[2] Andrews S, Hofmann T,Tsochantaridis I. Multiple instance learning with generalized support vector machines[C].//IEEE National Conference on Artificial Intelligence, Edmonton [s.n.],2002:943-944.

[3]Wang J,Zucker JD.Solving the multiple instance problem: a lazy learning approach [C].// IEEE the 17th International Conference on Machine Learning , San Francisco [s.n.] ,2000: 1119-1125.

[4]Zhi-Hua Zhou, Yu-Yin Sun, Yu-Feng Li. Multi-instance learning by treating instances as non-I.I.D.samples[C].//IEEE Proceedings of the 26th International Conference on Machine Learning ,Canada[s.n.], 2009: 1249-1256

[5]Rahmani R, Goldman SA. Missl: Multiple instance semi-supervised learning[C].//IEEE the 23rd International Conference on Machine Learning. Pittsburgh [s.n.], 2006:705–712.

[6] Z.-H Zhou ,J.-M Xu. On the relation between multi-instance learning and semi-supervised learning[C].//IEEE the 24th International Conference on Machine Learning. Corvalis, s.n.],2007:1167-1174.

[7]Z.-H. Zhou. Multi-instance learning from supervised view[J]. Journal of Computer Science and Technology, 2006, 21(5): 800-809.

[8] J.R.Quinlan. C4.5: programs for machine learning. Morgan Kaufmann, San Mateo, 1993..

[9] I. Witten and E. Frank. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2005.