Multiple-instance Learning based on Bernoulli Mixture Model

Lilin Yang*, Wei Wu
School of Wuhan University of Technology, Wuhan 230070, China
*Corresponding author e-mail: 409297468@qq.com

Abstract. Multiple-instance learning (MIL) is a form of weakly supervised learning. Its instances are arranged in groups (called bags), and labels are provided for the entire bag after training. The classic MIL method represents examples with pre-calculated features, and the classification process is also cumbersome. In this paper, we use neural networks to extract features, parameterize all transformations, and use Bernoulli mixture model to construct MIL models for baggage tags, using simpler network structures and more accurately solving these problems. Experiments show that our results can be competitive with the classical MIL algorithm on the MINST dataset.

1. Introduction

Supervised learning, semi-supervised learning, and unsupervised learning are several important methods in machine learning. They are all learning problems corresponding to a single sample and a label. We consider such a training set. The training set consists of a set of multiple-instance bags with classification labels, and each multiple-instance bag contains several instances without classification labels. Each sample package has its own mark. When the mark of a package is negative, the marks of all the samples in this package are negative; when the mark of a package is positive, at least one of the samples in this package has mark as positive. Such a classification problem is a Multiple instance learning (MIL) [1] problem.

MIL algorithms can be divided into three types based on example horizontal space, package horizontal space and embedded space. The first type of algorithm is to train an example-level classifier so that it can distinguish examples from positive and negative packets, and then use this classifier to judge each example in the test packet to get the judgment value of this packet. Common MIL algorithms based on example horizontal space include APR [1], MI-SVM [2], SMILE [3], SVR-SVM [4], and Clustering MIL [5]. The basic idea of the second type of algorithm is: first define a function that measures the distance between packets, and then embed the distance function in a standard distance-based classifier, such as SVM. Common multiple-example algorithms based on packet horizontal space include MI-SVM [2], Bayesian-kNN [6], Citation-kNN [6], MI-kernel [7], MIForest [8], and MinD [9]. The basic framework of the third type of algorithm is: each packet will be mapped to a single feature vector, which can describe the overall information related to the packet. Its representative algorithms are Simple-MI [10], DD-SVM [11], MILES [12], MILD [13], MILIS [14], and RSIS [15].

These classic methods represent instances by pre-calculated features, and do not use other methods for feature extraction. In general, there are differences between MIL methods at the instance level. These questions raise questions about the usability of the current MIL model to interpret the final
decision. In this paper, we propose a new method for extracting the global information of packets. This algorithm improves the interpretability of MIL, and also makes the network framework more concise and improves the accuracy of the algorithm. We use LeNet-5 [16] for feature extraction, construct a MIL model for luggage tags by Bernoulli Mixture Model, and train it by optimizing the log-likelihood function. We use neural networks to parameterize all transformations and train the model in an end-to-end manner by optimizing the unconstrained objective function. In the experiments, we proved that our algorithm has a certain performance improvement over the classic MIL algorithm on the MINST dataset.

2. Methodology

2.1. LeNet-5

![Figure 1. Architecture of LeNet-5.](image)

LeNet-5 has 7 layers. The input layer does not count the number of layers. Each layer has certain training parameters. Among them, the three convolution layers have more training parameters. Each layer has multiple filters, also called feature maps. Each filter extracts different pixel features from the output of the previous layer, and then each Feature Map has multiple neurons.

2.2. Bernoulli Mixture Model

Bernoulli distribution is also known as "zero-one distribution" and "two-point distribution". If the random variable $X$ only takes two values of 0 and 1, and the corresponding probability is:

$$\Pr(X = 1) = 1, \Pr(X = 0) = 1 - p, 0 < p < 1.$$  

(1)

The random variable $X$ is said to obey the Bernoulli distribution with parameter $p$. If $q = 1-p$, the probability function of $X$ can be written as:

$$f(x|p) = \begin{cases} p^x q^{1-x}, & x = 0, 1; \\ 0, & x \neq 0, 1. \end{cases}$$  

(2)

The Bernoulli mixture model can be regarded as a model composed of $K$ Bernoulli distributions. These $K$ sub-models are hidden variables of the mixed model. First, define a universal model of a mixed model consisting of $K$ sub-distributions in a D-dimensional space:

$$P(x|\Theta, \pi) = \sum_{k=1}^{K} \pi_k P(x|\theta_k)$$  

(3)
Among them \( \pi_k \) is the k-th sub-distribution mixing coefficient, also called weight, and satisfies \( \sum_{k=1}^{K} \pi_k = 1; \Theta = \{\theta_1, \theta_2, \ldots, \theta_K\} \) is a dataset containing N samples, and all the samples satisfy the same independent distribution. For mixture models, the maximum likelihood method is usually used to estimate the parameters \( \Theta \).

### 2.3. MIL based on Bernoulli Mixture Model

![Figure 2. The proposed approach with the Bernoulli Mixture Model as the MIL pooling. Red color corresponds to Bernoulli Mixture Model, blue color depicts a bag vector representation.](image)

The classical methods of MIL all represent instances by pre-calculated features, and do not use other methods for feature extraction. The method of extracting signs by convolutional neural network has made great contributions to various image processing methods, and LeNet-5 is a very efficient convolutional neural network for handwritten character recognition. We use LeNet-5 for feature extraction on our images.

Figure 2 is the MIL method proposed in this paper. The left convolution layer and the full convolution layer are part of LeNet-5. They perform feature extraction on the input picture, and then use the MIL constructed by the Bernoulli mixture model to perform the extracted features Judgment and classification, get the final classification result through the full convolutional neural network. MIL and deep learning methods allow the application of a flexible type of transformation, which can be trained end-to-end through back-propagation. Our example has two positive and negative labels, and the example package contains multiple examples, so the Bernoulli mixture model can build a MIL model for a bag label. The Bernoulli mixture model assigns different weights to multiple Bernoulli distributions, which also helps us find key examples.

### 3. Experiment

#### 3.1. Dataset

The MNIST dataset is from the National Institute of Standards and Technology (NIST). It consists of Training set images, Training set labels, Test set images, and Test set labels. The training dataset contains 60,000 samples, and the test dataset contains 10,000 samples. In the MNIST dataset, each sample is a 28 * 28 pixel handwritten digital picture, and each pixel is represented by a gray value.

#### 3.2. Training details

To create the test bag, we only sample images from the MNIST test set. During training, we only used the images in the MNIST training set. If a bag contains one or more images with a "9" label, the bag will have a front label. We chose "9" because it could easily be mistaken for "7" or "4". For all experiments, the LeNet5 model was used. We keep the default parameters of \( \beta_1 \) and \( \beta_2 \).

We use the area under the receiver operating characteristic curve (AUC) as our scoring standard. AUC is a model evaluation indicator in the field of machine learning. The AUC value is a probability value. When we randomly select a positive sample and a negative sample, the probability that the
current classification algorithm will rank this positive sample before the negative sample according to the calculated Score value is the AUC value. The larger the AUC value, the current classification. The more likely the algorithm is to rank positive samples before negative samples, the better the classification will be. In addition, we compare our method with the SVM-based MIL method (MI-SVM), which uses a Gaussian kernel on the original pixel features.

3.3. Compare results

| METHOD | 50   | 100  | 150  | 200  |
|--------|------|------|------|------|
| MI-SVM | 0.696| 0.852| 0.862| 0.898|
| BMM-MIL| 0.724| 0.872| 0.880| 0.901|

Table 2. The test AUC for MNIST-BAGS with on average 50 instances per bag for different numbers of training bags.

| METHOD | 50   | 100  | 150  | 200  |
|--------|------|------|------|------|
| MI-SVM | 0.824| 0.946| 0.959| 0.967|
| BMM-MIL| 0.843| 0.951| 0.960| 0.965|

In the experiments, we used different numbers of average bag sizes, 10 and 50, with a variance of 2,10. We also used a different number of examples in the package. Tables 1 and 2 are the AUC results tested on the MNIST dataset. Table 1 shows an average of 10 examples per packet, and Table 2 shows an average of 50 examples per packet, BMM-MIL represents our approach. After the same number of iterations of our algorithm and MI-SVM, the best results are selected for comparison, and we can find that our algorithm performs better, especially in the case of smaller bag sizes and fewer examples. When there are many examples, our results and MI-SVM are good.

4. Conclusion

In this paper, we propose a simple and accurate MIL algorithm, using the well-performing LeNet-5 to parameterize our neural network. In addition, we also propose the use of a Bernoulli mixture model to construct the MIL algorithm. Our experiments on the MINST handwritten data set show that our method has improved accuracy and works better in small samples, etc. This is important because the medical imaging problem contains only a few cases. We focus on simple MIL problems, but the Bernoulli mixture model can solve more complex problems, so we will want to apply it to multi-instance multi-label problems in the future, this problem is for further study.

References

[1] Thomas G. Dietterich, Richard H. Lathrop, Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles [J]. Artificial Intelligence, 1997, 89 (1).
[2] Andrews S, Tsochantaridis I, Hofmann T. support vector machines for multiple-instance learning [C] //Proceedings of the Advances in Neural Information Processing Systems, 2002, 561-568.
[3] Xiao Y, Liu B, Hao Z, et al. A Similarity-Based Classification Framework for Multiple-Instance Learning [J]. IEEE Transactions on Cybernetics, 2013, 44 (4): 500-515.
[4] Li F, Sminchisescu C. Convex multiple-instance learning by estimating likelihood ratio [C] //Advances in Neural Information Processing Systems. 2010: 1360-1368.
[5] Tax D M J, Hendriks E, Valstar M F, et al. The detection of concept frames using clustering multi-instance learning [C]/2010 20th International Conference on Pattern Recognition. IEEE, 2010: 2917-2920.
[6] Wang J, Zucker J D. Solving multiple-instance problem: A lazy learning approach [J]. 2000.
[7] Gärtner T, Flach P A, Kowalczyk A, et al. Multi-instance kernels[C]//ICML, 2002, 2 (3): 7.
[8] Leistner C, Saffari A, Bischof H. MIForests: Multiple-instance learning with randomized trees
[C] //European Conference on Computer Vision. Springer, Berlin, Heidelberg, 2010: 29-42.
[9] Cheplygina V, Tax D M J, Loog M. Multiple instance learning with bag dissimilarities [J].
Pattern Recognition, 2015, 48 (1): 264-275.
[10] Dong L. A comparison of multi-instance learning algorithms [D]. The University of Waikato,
2006.
[11] Chen Y, Wang J Z. Image categorization by learning and reasoning with regions [J]. Journal of
Machine Learning Research, 2004, 5 (Aug): 913-939.
[12] Chen Y, Bi J, Wang J Z. MILES: Multiple-instance learning via embedded instance selection [J].
IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006, 28 (12): 1931-1947.
[13] Li W J. MILD: Multiple-instance learning via disambiguation [J]. IEEE Transactions on
Knowledge and Data Engineering, 2009, 22 (1): 76-89.
[14] Fu Z, Robles-Kelly A, Zhou J. MILIS: Multiple instance learning with instance selection [J].
IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 33 (5): 958-977.
[15] Carbonneau M A, Granger E, Raymond A J, et al. Robust multiple-instance learning ensembles
using random subspace instance selection [J]. Pattern recognition, 2016, 58: 83-99.
[16] LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition
[J]. Proceedings of the IEEE, 1998, 86 (11): 2278-2324.