The Ninth International Symposium on Operations Research and Its Applications (ISORA'10)
Chengdu-Jiuzhaigou, China, August 19–23, 2010
Copyright © 2010 ORSC & APORC, pp. 160–167

Multiple Instance Learning via Multiple Kernel Learning[*]

Bing Yang   Qian Li   Ling Jing   Ling Zhen[†]

College of Science, China Agricultural University, Beijing 100083, P.R. China.

Abstract In this paper, we formulated a novel method to solve the classification problem within the multiple instance learning (MIL) contexts by multiple kernel learning. Despite the large number of SVM models, there are only a few models that can solve general MIL problems well. To improve the classification precision of SVM method with regard to MIL problem, this paper introduced multiple kernel learning method to the process of multiple instance learning, and proposed a new SVW model (MKMI-SVM), which based on the MI-SVM model. The solution for this model was presented, and some numerical experiments on benchmark data were taken into this paper too. Computational results on a number of datasets indicate that the proposed algorithm is competitive with other SVM methods.

Keywords Multiple instance learning, Support vector machines, multiple kernel learning, multi-class SVM

1 Introduction

Literatures [1–3] gave us an introduction about multiple instance classification problems. In this paper we gave a method to solve the problem which is mentioned in literature [10]. The problem to consist of *classifying positive and negative bags of points in the n-dimensional real space R^n on the following basis* is considered. A bag is classified as a positive bag if one or more instances in that bag are positive, otherwise it is classified as a negative bag.

This problem was firstly analyzed by T. G. Dietterich et al. [1] in the pharmic activity's prediction in the 90s, 20century. T. G. Dietterich considered every molecule as a bag in their analysis, and every low power shape represented an instance in the bag. This is the origin of multiple instance learning (MIL). The MIL problem has been existed for a long while; however, it is not a sudden result of pharmic activity's prediction. Previous machine learning [4, 5] didn't take this kind of problems' property into consideration formally, and the problem hasn't been exactly

[*] This work is supported by the Key Project of National Natural Science Foundation of China (No.10631070), the National Natural Science Foundation of China (No.10971223), the Chinese Universities Scientific Fund(No.2009-2-05), and the Graduate Basic Research Foundation of China Agricultural University Funded Project (No.15059208).

[†] Corresponding author. E-mail address: jingling_student@163.com.

defined until T. G. Dietterich's work.

Later the analysis of this problem aroused a number of Machine learning researchers' interest and a lot of research works have been done. Various methods for multiple instance classification problems have been proposed, including integer programming [6], expectation maximization [7], kernel formulations [8], and lazy learning [7]. Ray and Craven [9] provide an empirical comparison of several multiple instance classification algorithms and their non-multiple-instance counterparts. The classical SVM methods to solve MIL problem are MI-SVM and mi-SVM, which proposed by S. Andrews [6]. Based on their work, this paper added multiple kernel methods to the classical SVM methods, and gave a novel formulation for MIL problem. Meanwhile the strengths and the weakness about this new method have been discussed. Andrews et al. extend the single kernel SVM, while we begin with the multiple kernels SVM [10]. The use of the multiple kernels SVM allows us to get a better description of data's distributing as opposed to single kernel SVM. We include results in Sect. 4 which demonstrate that multiple kernel methods are much more computationally efficient and faster than classical methods.

The paper is organized as follows. In Sect. 2 we give a review of some interrelated concepts. In Sect. 3 we introduce our formulation of the multiple instance classification problems and state its properties. In Sect. 4 we present our numerical tests on five datasets. Section 5 concludes the paper.

2 The background about SVM method for multiple instance learning and multiple kernel learning

As the background of this paper, this section will introduce MI-SVM which is the classical SVM method for MIL problem and the standard multiple kernel SVM method.

2.1 Support vector machines for multiple instance learning

In this part, a brief review about classical SVM methods for multiple instance learning, MI-SVM and mi-SVM, will be shown. And we are going to introduce the basic idea, model and solving algorithm respectively.

Andrews et al. [6] have previously investigated extending support vector machines to the multiple instance classification problems using mixed-integer programming. They use integer variables either to select the class of points in positive bags or to identify one point in each positive bag as a "witness" point that must be placed on the positive side of the decision boundary. Each of these representations leads to a natural heuristic for approximately solving the resulting mixed-integer program.

S. Andrews gave an alternative way [6] of applying maximum margin ideas which is the main ideas of SVM to the MIL setting. They extend the notion of a margin from individual patterns to set of patterns. It is natural to define the functional margin of a bag with respect to a hyperplane by $\gamma_I = Y_{\mathrm{I}} \max_{i \in I}(\langle w, xi \rangle + b)$. Therefore based on the this notion of a bag margin, the SVM model has been redefined into

$$\text{MI-SVM} \qquad \begin{aligned} &\min_{w,b,\xi} && \frac{1}{2}\|w\|^2 + C\sum_I \xi_I \\ &s.t. && Y_I \max_{i\in I}(\langle w, x_i\rangle + b) \geq 1-\xi_I \qquad \forall I \\ & && \xi_I \geq 0 \end{aligned}$$

For solving this program, unfolding the max operations by introducing one inequality constraint for per instance has been done. For negative bags, the inequality constrains can be read as $-\langle w, xi\rangle - b \geq 1-\xi_I, \forall i \in I$, where $Y_I = -1$.

For positive bags, [6] introduces a selector variable $s(I) \in I$ which denotes the instance selected as the positive "witness" in per positive bag B_I. Meanwhile they gave two methods to select $s(I) \in I$ from B_I, MI-SVM and mi-SVM.

Both of MI-SVM and mi-SVM are casted as mixed-integer programs. They will be shown in algorithm1 and algorithm 2 in the following, respectively.

Algorithm1 mi-SVM algorithm

```
Initialize  $y_i = Y_I$  for  $i \in I$
  REPEAT
    Compute SVM solution w, b for data with imputed labels
    Compute outputs  $f_i = (W, X_i) + b$  for all  $X_i$  in positive bags
      Set  $y_i = \mathrm{sgn}(f_i)$  for every  $i \in I, Y_I = 1$
        FOR (every positive bag  $B_I$ )
          IF ( $\sum_{i\in I}(1+y_i)/2 == 0$ )
            Compute  $i^* = \arg\max_{i\in I} f_i$
            Set  $y_{i^*} = 1$
          END
        END
  WHILE (imputed labels have changed)
OUTPUT(w, b)
```

Algorithm2 MI-SVM algorithm

```
Initialize  $X_I = \sum_{i\in I} x_i / |I|$  for every positive bag  $B_I$
  REPEAT
    Compute QP solution w, b for data set with positive examples{ $X_I : Y_I = 1$ }
    Compute outputs  $f_i = (w, x_i) + b$  for all  $x_i$  in positive bags
    Set  $X_I = X_{s(I)}, s(I) = \arg\max_{i\in I} f_i$  for every  $I, Y_I = 1$
  WHILE ( selector variables  $s(I)$  have changed)
OUTPUT(w, b)
```

2.2 Multiple kernel learning

MI-SVM model by Andrews et al. extend the classical simple kernel SVM, while we begin with the multiple kernels SVM [11]. In this part, we will give an introduction about multiple kernels SVM's basic idea, model and main computing Algorithm.

Multiple kernels learning (MKL) aims at simultaneously learning a kernel and the associated predictor in supervised learning settings. Let $\{x_i, y_i\}_{i=1}^{l}$ is the learning set, where x_i belongs to some input space X and y_i is the target value for pattern x_i. For kernel algorithms in SVM, the solution of the learning problem is of the form

$$f(x) = \sum_{i=1}^{l} \alpha_i^* K(x, x_i) + b^*$$

where α_i^* and b^* are some coefficients to be learned from examples, while $K(\cdot,\cdot)$ is a given positive definite kernel associated with a reproducing kernel Hilbert space (RKHS) H.

In some situations, a learning practitioner may be interested in more flexible models. Recent applications have shown that using multiple kernels instead of a single one can enhance the interpretability of the decision function and improve performances [11]. In such cases, a convenient approach is to consider that the kernel $K(x,x')$ is actually a convex combination of basis kernels:

$$K(x,x') = \sum_{m=1}^{M} d_m K_m(x,x'),$$
$$with \quad d_m \geq 0, \quad \sum_{m=1}^{M} d_m = 1$$

where M is the total number of kernels. Each basis kernel K_m may either use the full set of variables describing x or subsets of variables stemming from different data sources [11]. Alternatively, the kernels K_m can simply be classical kernels (such as Gaussian kernels) with different parameters. Within this framework, the problem of data representation through the kernel is then transferred to the choice of weights d_m.

In the SVM methodology, the decision function is of the form $f(x) = \sum_{i=1}^{l} \alpha_i y_i \sum_{m=1}^{M} d_m K_m + b$, where the optimal parameters d_m, α_i and b are obtained by solving the dual of the following optimization problem [10]:

$$\begin{aligned} &\min_{\{f_m\},b,\xi,d} \frac{1}{2}\sum_m \frac{1}{d_m} \| f_m \|_{H_m}^2 + C\sum_i \xi_i \\ &s.t. \quad y_i \sum_m f_m(x_i) + y_i b \geq 1 - \xi_i \quad \forall i \\ &\qquad \xi_i \geq 0 \quad \forall i \\ &\qquad \sum_m d_m = 1, \quad d_m \geq 0 \quad \forall m \end{aligned}$$

The MKL formulation introduced by Bach et al. [12] and further developed by Sonnenburg et al. [13] consists in solving an optimization problem expressed above. Nowadays an effective method to solve this optimization problem is proposed by Alain Rakotomamonjy et al. in 2008 [11]. The main algorithm will be shown in algoithm3 in following.

Algorithm3 Simple MKL algorithm

```
Set $d_m = \frac{1}{M}$ for $m = 1,...,M$
  While stopping criterion not met do
  Compute $J(d)$ by using an SVM solver with $K = \sum_m d_m K_m$
  Compute $\frac{\partial J}{\partial d_m}$ for $m = 1,...M$ and descent direction $D(12)$ .
  Set $\mu = \arg\max_m d_m$ , $J^\dagger = 0$ , $d^\dagger = d$ , $D^\dagger = D$
      While $J^\dagger < J(d)$ do {descent direction update}
          $d = d^\dagger$ , $D = D^\dagger$
          $v = \arg\min_{\{m|D_m<0\}} -d_m / D_m, \gamma_{\max} = -d_v / D_v$
          $d^\dagger = d + \gamma_{\max} D$ , $D_\mu^\dagger = D_\mu - D_v$ , $D_v^\dagger = 0$
          Compute $J^\dagger$ by using an SVM solver with $K = \sum_m d_m^\dagger K_m$
      End while
   Line search along $D$ for $\gamma \in [0, \gamma_{\max}]$ {calls an SVM solver for each $\gamma$ trial value}
  $d \leftarrow d + \gamma D$
  End while
```

3 MKMI-SVM Classification Model and Algorithm

Multiple kernel SVM is used for some situations where a machine learning practitioner may be interested in more flexible models. We can expect multiple kernel learning will has a better performance in MIL problem for two reasons. Obviously, since the highly complicated description about real object in MIL that the special problem, a flexible model is necessary for the learning task. Meanwhile the enhance about the interpretability of the decision function, more effectible computation and higher predication accuracy not only can be expected, but also are our hopes in MIL. Therefore, it is significant to add multiple kernel method to MIL problem. In this section, the model and algorithm of MKMI-SVM will be given.

In MKMI-SVM method, we also defined the functional margin of a bag with respect to a hyperplane by $\gamma_I = Y_{\rm I} \max_{i\in I}(\langle w, xi\rangle + b)$. Based on this rules, the inequality constraints in multiple kernel SVM can be changed for solving MIL problem. Therefore the MKMI-SVM model can be expressed a new optimization problem showed in following:

$$
\text{MKMI-SVM}\qquad
\begin{aligned}
&\min_{\{f_m\},b,\xi,d} \frac{1}{2}\sum_m \frac{1}{d_m}\|f_m\|_{H_m}^2 + C\sum_I \xi_I \\
&s.t. \quad Y_I \max_{i\in I}(\sum_m f_m(x_i) + b) \ge 1 - \xi_I \qquad \forall I \\
&\qquad\quad \xi_I \ge 0 \\
&\qquad\quad \sum_m d_m = 1, \qquad d_m \ge 0 \quad \forall m
\end{aligned}
$$

Since the first constraint in our multiple instance formulation contains the max operations. We also unfolded this max operation as [6]. For negative bags, the inequality constraint can be read as $-\langle w, xi\rangle - b \ge 1 - \xi_I, \forall i \in I$, where $Y_I = -1$. For positive bags, a selector variable $s(I) \in I$ which denotes the instance selected as the positive instance in per positive bag B_I will be gave. For d_m , $s(I) \in I$ and α, b , alternately compute one set of variables when hold other sets. This leads to the successive solution of MKMI-SVM programs that underly our algorithm which we specify now.

Algorithm4 MKMI-SVM Algorithm

```
Initialize $y_i = Y_i$ for $i \in I$
REPEAT
Compute MK-SVM solution $K = \sum_{m=1}^{M} d_m K_m(x_n, x_i), \alpha, b$ for data set with imputed labels
Compute outputs $f_i = \sum_{n=1}^{l} \alpha_i (\sum_{m=1}^{M} d_m K_m(x_n, x_i)) + b$ for all $x_i$ in positive bags
FOR (every positive bag $B_I$ )
    IF ( $\sum_{i\in I} (1 + y_i)/2 == 0$ )
    Compute $i^* = \arg\max_{i\in I} f_i$
    Set $y_{i*} = 1$
    END
END
WHILE (imputed labels have changed)
OUTOUT ( $d_m, \alpha, b$ )
```

In practice, computing the MKMI-SVM Algorithm 4 may be faster than classical MI-SVM when you should change your kernel to check more kernel Hilbert space (RKHS). If the number of kernel you should compute is N, the classical SVM methods will compute their Algorithm 5 times. And the MKMI-SVM' time is equal to its number of iterations time. In [11], a lot of numerical testing had been done to compare which is faster. Multiple kernel learning often has the better performances.

4 Numerical Experiments

In this section, some numerical experiments will be done for testing the MKMI-SVM's capabilities in MIL problem. To evaluate the capabilities of MKMI-SVM method, we have performed some experiments on benchmark data. In this paper, we reported results on 5 datasets, two from the UCI machine learning repository [14], and three from [6]. Detailed information about these datasets is summarized in Table 1. We use the datasets from [6] to evaluate our multiple kernel classification algorithms. These three datasets are from an image annotation task in which the goal is to determine whether or not a given animal is present in an image. The two datasets from the UCI repository [14] are the Musk datasets, which are commonly used in multiple instance classification.

Table 1 Description of the datasets used in the experiments. Elephant, Fox and Tiger datasets are used in [6], while Musk-1 and Musk-2 are available from [14]. +Bags denotes the number of positive bags in each dataset, while +Instances denotes the total number of instances in all the positive bags. Similarly, −Bags and −Instances denote corresponding quantities for the negative bags

Data set	+bag	+instances	-bag	-instances	features
Elephant	100	762	100	629	143
Fox	100	647	100	673	143
Tiger	100	544	100	676	143
Musk-1	47	207	45	269	166
Musk-2	39	1017	63	5581	166

We compare our multiple kernels classification algorithm to the mi-SVM and MI-SVM [6] on three image datasets. Since Andrews et al. also report results on Zhang and Goldman's expectation maximization approach EM-DD [7] on these datasets [6]; we include those results here as well. Table 2 reports results comparing MKMI-SVM to mi-SVM, MI-SVM and EM-DD. Accuracy results for mi-SVM, MI-SVM and EM-DD were taken from [6]. Accuracy for MIMK-SVM was measured by averaging ten ten-fold cross validation runs. The multiple kernels for MKMI-SVM were selected by 10 different Gaussian kernel, kernel parameters form 2^{-5} to 2^{4}. The parameters C for MKMI-SVM were selected from the set $\{2^{i} \mid i = -5, . . . , 4\}$ by ten-fold cross validation on each training samples of the image datasets.

Table 2 MKMI-SVM, mi-SVM [6], MI-SVM [6] and EM-DD [7] testing accuracy used averaged over ten ten-fold cross validation experiments. The datasets are those used by Andrews et al. in [6]. Best accuracy on each dataset is in bold.

Data set	MKMI-SVM	mi-SVM	MI-SVM	EM-DD
Elephant	81.8%	**82.2%**	81.4%	78.3%
Fox	**58.7%**	58.2%	57.8%	56.1%
Tiger	**84.0%**	78.4%	**84.0%**	72.1%

In order to evaluate the difference between the algorithms more precisely, we used the Friedman test [17] on the results reported in Table 2. The Friedman test is a nonparametric test that compares the average ranks of the algorithms, where the algorithm with the highest accuracy on a dataset is given a rank of 1 on that dataset, and the algorithm with the worst accuracy is given a rank of 5. Therefore the average rank was 1.3 for MKMI-SVM, 1.5 for mi-SVM, 2.3 for MI-SVM, and 4 for EM-DD. The better performance by MKMI-SVM expressed on MKL problem was clearly showed.

Table 3 MKMI-SVM, mi-SVM [6], MI-SVM [6], EM-DD [7], DD [15], MI-NN [16], IAPR [1], and MIK [8] ten-fold testing accuracy on the Musk-1 and Musk-2 datasets. Best accuracy is in bold.

Dataset	MKMI -SVM	mi-SVM	MI-SVM	EM-DD	DD	MI-NN	IAPR	MIK
Musk-1	88.6%	87.4%	77.9%	84.8%	88.0%	88.9%	**92.4%**	91.6%
Musk-2	85.2%	83.6%	84.3%	84.9%	84.0%	82.5%	**89.2%**	88.0%

Table 3 gives ten-fold cross validation accuracy results for MKMI-SVM using the same test method on the Musk-1 and Musk-2 datasets which are available from the UCI repository [14]. In table 3, we can see that MKMI-SVM got the best accuracy in all SVM methods, but some simple method like IAPR methods, got the better result on the contrary. It showed that MKMI-SVM will just have a better performance on the complicated dataset of which the acting feature and its reciprocity in classification is not very clear, but for the ordinary datasets of which the acting feature and its reciprocity is clearly enough, it is proper to perform some simple methods. Obviously, this is in line with the principle of Occam's razor; meanwhile it can suggest us the type of dataset for which MKMI-SVM method is proper to be used.

5 Conclusion and Outlook

This paper has introduced a mathematical programming formulation of the multiple instance problems that has used multiple kernel learning. Results on previously published datasets indicate that our approach is effective at some situation where a machine learning practitioner may be interested in more flexible models. Furthermore, multiple kernels learning often cost less than simple kernel for learning in multiple kernel Hilbert space, and computing the MKMI-SVM maybe faster than classical simple kernel method in practice. Improvements in the mathematical programming formulation and evaluation using a wide variety of datasets and algorithms, such as those in [17], are promising avenues of future research.

References

[1] Dietterich, T.G., Lathrop, R.H., Lozano-Perez, T.: Solving the multiple-instance problem with axis-parallel rectangles. Artif. Intell. 89, 31–71 (1998)

[2] Auer, P.: On learning from multi-instance examples: empirical evaluation of a theoretical approach. In: Proceedings of 14th International Conference on Machine Learning, pp. 21–29. Morgan Kaufmann, San Mateo (1997)

[3] Long, P.M., Tan, L.: PAC learning axis aligned rectangles with respect to product distributions from multiple instance examples. Mach. Learn. 30(1), 7–22 (1998)

[4] Friedman J H, Stuetzle W. Projection pursuit regression. Journal of the American Statistical Association, 1981, 76(376): 817-823.

[5] Lindsay R, Buchanan B, Feigenbaum E, Lederberg J. Applications of Artificial Intelligence to Organic Chemistry: The Dendral Project, New York, NY: McGraw-Hill, 1980.

[6] Andrews, S., Tsochantaridis, I., Hofmann, T.: Support vector machines for multiple-instance learning. In: Becker, S., Thrun, S., Obermayer, K. (eds.) Advances in Neural Information Processing Systems 15, pp. 561–568. MIT Press, Cambridge (2003)

[7] Zhang, Q., Goldman, S.A.: EM-DD: an improved multiple-instance learning technique. In: Neural Information Processing Systems 2001, pp. 1073–1080. MIT Press, Cambridge (2002)

[8] Gartner, T., Flach, P.A., Kowalczyk, A., Smola, A.J.: Multi-instance kernels. In: Sammut, C., Hoffmann, A. (eds.) Proceedings of 19th International Conference on Machine Learning, pp. 179–186. Morgan Kaufmann, San Mateo (2002)

[9] Ray, S., Craven, M.: Supervised versus multiple instance learning: an empirical comparison. In: Proceedings of 22nd International Conference on Machine Learning, Bonn, Germany, vol. 119, pp. 697–704. Assoc. Comput. Mach., New York (2005)

[10] O.L. Mangasarian, E.W. Wild: Multiple Instance Classification via Successive Linear Programming. In: J Optim Theory Appl (2008) 137: 555–568

[11] Alain Rakotomamonjy, Francis R. Bach, St ́ephane Canu, Yves Grandvalet: SimpleMKL. In: Journal of Machine Learning Research 9 (2008) 2491-2521

[12] F. Bach, G. Lanckriet, and M. Jordan: Multiple kernel learning, conic duality, and the SMO algorithm. In: Proceedings of the 21st International Conference on Machine Learning, pages 41–48, 2004a.

[13] S. Sonnenburg, G. Ratsch, C. Schafer, and B. Scholkopf.: Large scale multiple kernel learning. In: Journal of Machine Learning Research, 7(1):1531–1565, 2006.

[14] Murphy, P.M., Aha, D.W.: UCI Machine Learning Repository (1992). www.ics.uci.edu/~mlearn/MLRepository.html

[15] Maron, O., Ratan, A.L.: Multiple-instance learning for natural scene classification. In: 15th International Conference on Machine Learning, San Francisco, CA. Morgan Kaufmann, San Mateo(1998)

[16] Ramon, J., De Raedt, L.: Multi-instance neural networks, In: Proceedings of ICML-2000. Workshop on Attribute-Value and Relational Learning (2000)

[17] Ray, S., Craven, M.: Supervised versus multiple instance learning: an empirical comparison. In: Proceedings of 22nd International Conference on Machine Learning, Bonn, Germany, vol. 119, pp. 697–704. Assoc. Comput. Mach., New York (2005)