IEICE TRANS. INF. & SYST., VOL.E94–D, NO.2 FEBRUARY 2011

LETTER

Laplacian Support Vector Machines with Multi-Kernel Learning

Lihua GUO[†a)], *Nonmember and* Lianwen JIN[†], *Member*

SUMMARY The Laplacian support vector machine (LSVM) is a semi-supervised framework that uses manifold regularization for learning from labeled and unlabeled data. However, the optimal kernel parameters of LSVM are difficult to obtain. In this paper, we propose a multi-kernel LSVM (MK-LSVM) method using multi-kernel learning formulations in combination with the LSVM. Our learning formulations assume that a set of base kernels are grouped, and employ l_2 norm regularization for automatically seeking the optimal linear combination of base kernels. Experimental testing reveals that our method achieves better performance than the LSVM alone using synthetic data, the UCI Machine Learning Repository, and the Caltech database of Generic Object Classification.

key words: *semi-supervised learning, manifold regularization, multi-kernel learning, Laplacian support vector machine*

1. Introductions

Supervised learning algorithms require a large amount of labeled data, which is often difficult or costly to obtain. Semi-supervised methods offer an interesting solution for this requirement, enable learning from both labeled and unlabeled data. In semi-supervised classification, the ultimate goal is to find a classifier that not only minimizes classification errors with labeled examples, but also can be compatible with the input distribution by monitoring values on unlabeled points [1]–[3]. Based on different problem settings, semi-supervised methods can be classified into two main categories: transductive learning (TL) [4] and semi-supervised inductive learning (SSIL) [5]. Recent studies have revealed that the success of SSIL depends on certain semi-supervised assumptions about the distribution of the data [6], such as the manifold assumption, which utilizes the fact that the distribution of the data has a low dimensional manifold. The underlying geometry of the data can typically be captured by representing the data as a graph, with samples as the vertices, and pair-wise similarities between the samples as edge weights. Based on this assumption, several graph-based algorithms including the label propagation [6], Markov random walk [7], graph cut [8], spectral graph transducer [9], and low-density separation [10] algorithms have been proposed in the literature. Recently, Belkin [11] proposed a version of the Laplacian support vector machine (LSVM) that used manifold regularization for inductive learning by constructing a maximum-margin classifier and penalizing the corresponding inconsistency with a similarity matrix. The LSVM used kernel functions, which operated in the feature space without ever computing the coordinates of the data, but rather by simply computing the inner products between all pairs of data. There are numerous forms of kernel functions in common use, such as the Linear, Poly and RBF so on. Since the feature space of data is not known, it should be based on empirical considerations, i.e. what kind of kernel parameters can be successfully presented for the feature space of data, and the optimal parameters, such the types and parameters of kernel, will be difficult to obtain. Kernel parameters selection is conventionally performed through repeated cross validation over a range of kernels and their parameters. Unfortunately, the LSVM model is one of semi-supervised learning framework, the unlabeled data cooperate with the labeled data for learning together and the cross validation is not used, so the LSVM still face the problems of kernel selection.

Multi-kernel learning (MKL) is an attractive tool for tracking many supervised learning tasks [12]–[14], and MKL algorithms have achieved very good results with challenging real-world applications [15], [16]. Inspired by these previous findings, our paper develops a new LSVM model involving multi-kernel learning (MK-LSVM) to automatically seek optimal parameters of kernel. Firstly, we anew model the semi-supervised learning problem based on the LSVM framework of Belkin [11], and change the single kernel learning into multi kernel learning; secondly, in the semi-supervision framework, we add the L2 norm regulation; finally, we use the Newton's descent method and iterate optimizations to get the direct solution.

2. MK-LSVM Algorithm

Belkin [11] proposed an LSVM classifier based on the manifold regularization, which extended the SVM by solving the following problem:

$$\min_{f\in H_K} \frac{1}{l}\sum_{i=1}^{l}(y_i - f(x_i))_+ + \gamma_A \|f\|_K^2 + \frac{\gamma_I}{(u+l)^2} f^T L f \quad (1)$$

The solution to the above problem is given by:

$$f^*(x) = \sum_{i=1}^{l+u} a_i^* K(x, x_i) \quad (2)$$

The kernel trick is a useful tool for mapping low-

Manuscript received July 14, 2010.
Manuscript revised September 27, 2010.

[†]The authors are with School of Electronic and Information Engineering, South China University of Technology, Guangzhou, 510641, China.
a) E-mail: guolihua@scut.edu.cn

DOI: 10.1587/transinf.E94.D.379

dimensional datasets to a higher number of dimensions for seeking the optimal classifying plane, but defining the optimal parameters of kernel function is difficult, based on the multi-kernel learning, the linear combination of given base kernels may be the optimal solution. Inspired by that, we introduce the multi-kernel learning into the LSVM, and add norm regularization of kernel coefficients in the semi-supervised learning framework. The L1 norm regularization has been more popular because of sparse kernel mixture, but the Lp for P>1 outperforms L1 MKL due to non-sparsity in the kernel weight which avoids over-fitting in some situations [17], [18]. In our method, L2 norm regularization is used for consideration that the system's optimization is easy.

Support we have a set of base kernels $K = \left\{k_j(x,x)\right\}, j = 1, 2, \ldots, m$. In the laplacian support vector machine with multi-kernel learning, we anew describe the semi-supervised learning problem as:

$$\begin{aligned}\min_{f\in H_K} \frac{1}{l}\sum_{i=1}^{l}(y_i - f(x_i))_+ + \gamma_A \|f\|_K^2 \\ + \frac{\gamma_I}{(u+l)^2} f^T L f + \gamma_\sigma \|D\|^2\end{aligned} \tag{3}$$

The solution is given by:

$$f^*(x) = \sum_{i=1}^{l+u} a_i^* \sum_{j=1}^{m} d_j K_j(x, x_i) \tag{4}$$

Where $D = \left\{d_j\right\} j = 1, 2, \ldots, m$ is the coefficient of the linear combination of base kernels. $\|f\|_K^2$ is the regulation of the classifying function, which is $a^T DKa$. $f^T Lf$ is the manifold regulation, L is the Laplacian matrix, $\|D\|^2$ is the l_2 norm regularization of coefficients, γ_A, γ_I and γ_σ are the coefficients, which are used to balance the loss function and three regularizations in function space. The optimal values of these coefficients are always practical ones and set according to some empirical consideration. Our method focuses on how to seek the optimal coefficients of base kernels.

Often in SVM formulations, an un-regularized bias term b is added to the above form. Again, the primal problem can be easily described as the following:

$$\begin{aligned}\min_{a\in R^{l+u}, \xi\in R^l, d\in R^m} \frac{1}{l}\sum_{i=1}^{l}\xi_i + \gamma_A a^T DKa \\ + \frac{\gamma_I}{(u+l)^2} a^T (DK)^T L(DK)a + \gamma_\sigma D^T D\end{aligned} \tag{5}$$

subject to:

$$y_i\left(\sum_{k=1}^{l+u} a_k \sum_{j=1}^{m} d_j K_j(x_i, x_k) + b\right) \geq 1 - \xi_i, i = 1, \ldots, l$$

$$\xi_i \geq 0$$

$$i = 1, \ldots, l$$

Introducing the lagrangian:

$$\begin{aligned}\Omega(a,\xi,b,\beta,\varsigma,d) = \frac{1}{l}\sum_{i=1}^{l}\xi_i + \gamma_A a^T DKa \\ + \frac{\gamma_I}{(u+l)^2} a^T (DK)^T L(DK)a + \gamma_\sigma D^T D \\ - \sum_{i=1}^{l}\beta_i\left(y_i\left(\sum_{k=1}^{l+u} a_k \sum_{j=1}^{m} d_j K_j(x_i, x_k) + b\right) - 1 + \xi_i\right) \\ - \sum_{i=1}^{l}\varsigma_i\xi_i\end{aligned} \tag{6}$$

The dual requires the following steps:

$$\frac{\partial\Omega}{\partial b} = 0 \Rightarrow \sum_{i=1}^{l}\beta_i y_i = 0$$

$$\frac{\partial\Omega}{\partial\zeta_i} = 0 \Rightarrow \frac{1}{l} - \beta_i - \varsigma_i = 0$$

using above identities, we formulate a reduced lagrangian as:

$$\begin{aligned}\Omega(a,\xi,b,\beta,\varsigma,d) = a^T(\gamma_A DK \\ + \frac{\gamma_I}{(u+l)^2}(DK)^T L(DK))a + \gamma_\sigma D^T D \\ - a^T DKJ^T YB + \sum_{i=1}^{l}\beta_i\end{aligned} \tag{7}$$

where $J = [I, 0]_{l\times l+u}$ is a $l \times (l+u)$ matrix with I as $l \times l$ identity matrix, $Y = diag(y_1, y_2, \ldots y_l)$ and $B = [\beta_i]$. Taking derivative of the reduced lagrangian with respect to α:

$$\begin{aligned}\frac{\partial\Omega(a,\xi,b,\beta,\varsigma,d)}{\partial a} = 0 \Rightarrow \\ a = \frac{1}{2}\left(\gamma_A DK + \frac{\gamma_I}{(u+l)^2}(DK)^T L(DK)\right)^{-1} DKJ^T YB\end{aligned} \tag{8}$$

Define $P = \gamma_A I + \frac{\gamma_I}{(u+l)^2}(DK)^T L$, It implies:

$$a = \frac{1}{2}P^{-1}J^T YB \tag{9}$$

$$\begin{aligned}\Omega(a,\xi,b,\beta,\varsigma,d) \\ = \sum_{i=1}^{l}\beta_i - \frac{1}{2}B^T\frac{1}{2}(P^{-1}J^T Y)^T DKJ^T YB + \gamma_\sigma D^T D\end{aligned} \tag{10}$$

If fixing D, and define $Q = \frac{1}{2}(P^{-1}J^T Y)^T DKJ^T Y$, then semi-supervised learning problem can be:

$$\min \Omega(a,\xi,b,\beta,\varsigma,d) \equiv \min\left(\sum_{i=1}^{l}\beta_i - \frac{1}{2}B^T QB\right) \tag{11}$$

This function can be implemented using a standard SVM solver with the quadratic form induced by the above matrix. After solving this problem, the D can be updated. Taking the derivative of the reduced Lagrangian with respect to D:

Table 1 The proposed MK-LSVM algorithm.

```
MK-LSVM
Input: $l$ labeled examples $\{(x_i, y_i)\}_{i=1}^{l}$, $u$ unlabeled examples $\{(x_i, y_i)\}_{i=l+1}^{l+u}$
Output: Estimated function $f : \Re^n \rightarrow \Re$
Step1: Choose $\gamma_A, \gamma_I, \gamma_\sigma$, and randomly set the initial coefficient $D = \{d_j\}$
Step2: Compute the gram matrix $DK = \sum_{p=1}^{m} d_p K_p(x_i, x_j)$, where $K_p(x_i, x_j)$ is the base kernel matrix.
Step3: Compute graph laplacian matrix:$L = P - W$, where $W$ is the weights of data adjacency graph, and $P$ is a diagonal matrix given by $P_{ii} = \sum_{j=1}^{l+u} W_{i,j}$.
Step4: Compute $\alpha^*$ using Equations 11 and 9 together with the SVM QP solver for soft margin loss.
Step5: Update the $D$ using Equations 14 until $D_{new} - D_{old} < \varepsilon$.
Step6: If the loss function in Equations 3 doesn't decrease, then exit, otherwise go to step 2
```

$$
\begin{aligned}
&\frac{\partial\Omega(a,\xi,b,\beta,\varsigma,d)}{\partial D} = \\
&\frac{1}{4}\frac{\partial B^T((\gamma_A I + \frac{\gamma_I}{(u+l)^2}(DK)^T L)^{-1} J^T Y)^T DKJ^T YB}{\partial D} \\
&+2\gamma_\sigma D
\end{aligned}
\tag{12}
$$

The function can be simplified as:

$$
\frac{\partial\Omega(a,\xi,b,\beta,\varsigma,d)}{\partial D} = \frac{1}{4}\gamma_A B^T Y^T JKJ^T Y P^{-2} B + 2\gamma_\sigma D \tag{13}
$$

Using the Newton descend method,

$$
\begin{aligned}
D_{new} &= D_{old} + \delta\frac{\Omega}{\partial\Omega/\partial D} \\
&= D_{old} + \delta\frac{4\Omega}{\gamma_A B^T Y^T JKJ^T Y P^{-2} B + 8\gamma_\sigma D}
\end{aligned}
\tag{14}
$$

The pseudo code of MK-LSVM is summarized in Table 1.

3. Experimental Results

Experiment 1: Two synthetic datasets, i.e. two-circle and star, are chosen as the test datasets, and labeled examples are marked using red and blue colors. The results of the LSVM method [11] are shown in Fig. 1 to Fig. 4. Kernel functions are RBF ($K(x, x_i) = \exp\{\frac{\|x-x_i\|^2}{2\sigma^2}\}$), with parameter σ of 0.35 and 0.1, respectively. When using RBF function (σ=0.35), the LSVM can perfectly classify the star dataset, but fails with the two-circle dataset. However, when using the RBF function (σ =0.1), the two-circle dataset can be successfully classified, but the star dataset cannot. This result confirms that the optimal parameters of kernel function are difficult to obtain. Our MK-LSVM method chooses a set of base kernels, such as Linear kernels ($K(x, x_i) = x \cdot x_i$), Poly kernels ($K(x, x_i) = \|1 + xx_i\|^t$) and RBF kernels ($K(x, x_i) = \exp\{\frac{\|x-x_i\|^2}{2r^2}\}$). The kernel parameters of Poly are $t_{poly} \in (2, 3, 4, 5, 6)$, and the kernel parameters of RBF are $\sigma_{RBF} \in (0.01, 0.05, 0.1, 0.2, 0.4, 0.8, 1.6, 3.2, 6.4, 12.8)$.

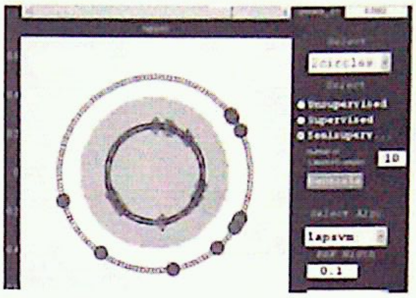


Fig. 1 Results of the two circles dataset using the LSVM [11] with RBF kernels with a width of 0.1.

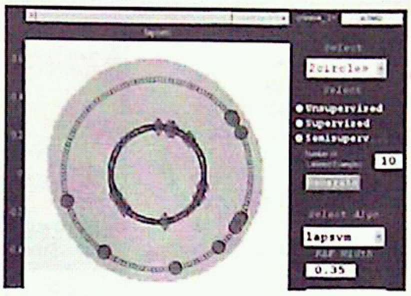


Fig. 2 Results of the two circles dataset using the LSVM [11] with RBF kernels with a width of 0.35.

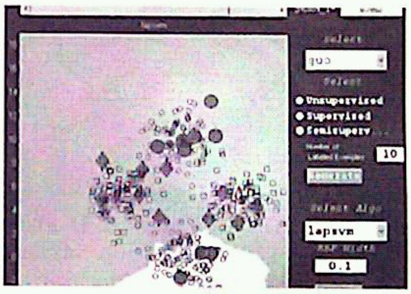


Fig. 3 Results of the star dataset using the LSVM [11] with RBF kernels with a width of 0.1.

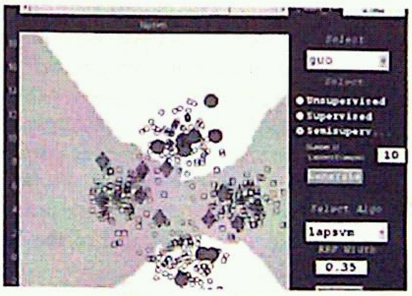
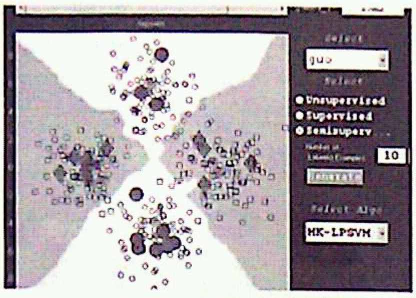


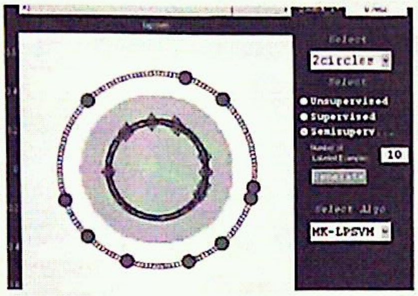
Fig. 4 Results of the star dataset using the LSVM [11] with RBF kernels with a width of 0.35.

Since our method can automatically seek the optimal parameters, the two synthetic datasets can be successfully classified without the need to define the kernel parameters in ad-



Select
guo
Select
● Unsupervised
● Supervised
○ Semisuperv...
Number of Labeled Examples 10
Select Algo
MK-LPSVM

Fig. 5 Results of the star synthetic dataset using MK-LSVM.



Select
2circles
Select
● Unsupervised
● Supervised
○ Semisuperv...
Number of Labeled Examples 10
Select Algo
MK-LPSVM

Fig. 6 Results of the two-circle synthetic dataset using MK-LSVM.

Table 2 Three experiments: one-versus-rest multi-class average error rates.

method	SVM [20]	LSVM [11]	MK-LSVM (our method)
USPS (error)	23.60	12.67	**12.50**
Wine (error)	17.08	8.86	**6.96**
TEXT (error)	19.24	10.41	**5.46**

vance, as shown in Fig. 5 and Fig. 6.

Experiment 2: The UCI Machine Learning Repository [19] (USPS, Wine) and TEXT categorization (available at: http://vikas.sindhwani.org/manifoldregularization.html) databases are used in this experiment. Comparisons are made using inductive methods (SVM) [20] and LSVM [11]. For simulating the cross validation for overcoming the overfitting of model, we design one-vs-rest multi-class experiments on USPS data with l=50 and u=1957 with 10 random splits, the wine data with l=20 and u=158 with 10 random splits, and the TEXT data with l=50 and u=1896 with 10 random splits. Each split is tested independently, and calculated the average performance. The average performance of the different methods is shown in Table 2. From Table 2, it can be seen that the average error rate of our proposed MK-LSVM method is much lower than that of the LSVM and SVM after testing ten splits.

Experiment 3: Generic Object Classification is a challenging topic in computer vision and machine learning. To confirm the validity of our proposed algorithm, we use Caltech-5 datasets [21], which include five objects, such as plane, car, background, leaves and faces. The PHOW features [22] are extracted from the datasets. Comparisons are made with inductive methods (SVM) [20], regularized least squares (RLS) [11] and LSVM [11]. Figure 7 shows the performance (error rate) of inductive and semi-supervised classifiers on unlabeled and test sets as a function of the number of labeled examples in the training set. The benefit of unlabeled data can be determined by comparing the performance curves of inductive and semi-supervised classifiers, revealing that our proposed MK-LSVM method achieves the highest performance.

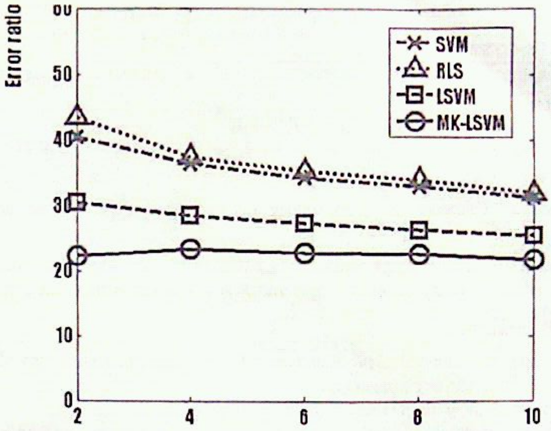


Fig. 7 Performance of inductive and semi-supervised classifiers.

4. Conclusion

The conventional LSVM method is one of the most important semi-supervised learning methods in machine learning applications. However, the optimal kernel parameters of LSVM are difficult to define. Based on the notion that the linear combination of given base kernels can provide optimal solutions, the present paper proposes an LSVM method with multi-kernel learning based on manifold regularization. Because multi-kernel learning can automatically determine the linear combination of base kernels for adapting all datasets and applications, our method is able to solve the shortcomings of LSVM kernel selection, which is the main essential advantage of our method. We test our MK-LSVM method using synthetic data, UCI Machine Learning Repository data and Caltech-5 datasets. The results have revealed that our method can efficiently solve a semi-supervised learning problem in the absence of training datasets.

In multi-kernel learning, there are different norm regularizations. The L1 norm regularization is more popular in previous studies because it outputs sparse kernel mixture, but the Lp for P>1 maybe outperforms L1 norm regularization due to non-sparsity of the kernel weight in some applications. The selection of best norm regularization is a very interesting research topic in semi-supervised learning framework that merits our future study.

Acknowledgments

This work is supported in part by the National Science Foundation of China (NSFC) (Grant no.60902087 and 61075021) and the Fundamental Research Funds for the Central Universities of SCUT (Grant no.2009ZM0036, 2009ZM0059 and 2009ZZ0014).

References

[1] M. Seeger, Learning with labeled and unlabeled data, Technical Report, School of Informatics, The University of Edinburgh, Edinburgh, U.K., 2000.

[2] X. Zhu, Semi-supervised learning literature survey, Technical Report TR-1530, Department of Computer Science, University of Wisconsin, Madison, U.S.A., 2005.

[3] O. Chapelle, B. Scholkopf, and A. Zien, Semi-Supervised Learning, Cambridge, MIT Press, MA, 2006.

[4] V.N. Vapnik, Statistical Learning Theory, Wiley, New York, 1998.

[5] G. Haffari, A survey on inductive semi-supervised learning, Technical Report, Department of Computer Science, Simon Fraser University, Vancouver, Canada, 2006.

[6] X. Zhu and Z. Ghahramani, Learning from Labeled and Unlabeled Data with Label Propagation, Technical Report CMU-CALD-02-107, Carnegie Mellon Univ., 2002.

[7] M. Szummer and T. Jaakkola, "Partially labeled classification with Markov random walks," Proc. Neural Information Processing Systems Conf., pp.945–952, 2002.

[8] A. Blum and S. Chawla, "Learning from labeled and unlabeled data using graph min-cuts," Proc. 18th Int'l Conf. Machine Learning, pp.19–26, 2001.

[9] T. Joachims, "Transductive learning via spectral graph partitioning," Proc. 20th Int'l Conf. Machine Learning, pp.290–297, 2003.

[10] O. Chapelle and A. Zien, "Semi-supervised classification by low density separation," Proc. 10th Int'l Workshop Artificial Intelligence and Statistics, pp.57–64, 2005.

[11] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," Journal of Machine Learning Research, vol.7, pp.2399–2434, 2006.

[12] S. Sonnenburg, G. Raetsch, C. Schaefer, and B. Schoelkopf, "Large scale multiple kernel learning," Journal of Machine Learning Research, vol.7, pp.1531–1565, 2006.

[13] F.R. Bach, "Exploring large feature spaces with hierarchical multiple kernel learning," NIPS, pp.105–112, 2008.

[14] A. Rakotomamonjy, F. Bach, S. Canu, and Y. Grandvalet, "Simplemkl," Journal of Machine Learning Research, vol.9, pp.2491–2521, 2008.

[15] M. Varma and D. Ray, "Learning the discriminative power invariance trade-off," Proc. International Conference on Computer Vision, pp.1–8, 2007.

[16] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," Proc. 6th Indian Conference on Computer Vision, Graphics and Image Processing, pp.722–729, 2008.

[17] M. Kloft, U. Brefeld, S. Sonnenburg, P. Laskov, K.-R. Muller, and A. Zien, "Efficient and accurate lp-norm multiple kernel learning," Advances in Neural Information Processing Systems, vol.22, pp.997–1005, MIT Press, 2009.

[18] R. Tomioka and T. Suzuki, "Sparsity-accuracy trade-off in MKL," NIPS 2009 Workshop, Whistler, Canada, 2009.

[19] UCI machine learning repository, http://www.ics.uci.edu/ mlearn/

[20] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," Software available at http://www.csie.ntu.edu.tw/ ˜cjlin/libsvm, 2001.

[21] R. Fergus, P. Perona, and A. Zisserman, "Object class recognition by unsupervised scale-invariant learning," Proc. CVPR, pp.264–271, 2003.

[22] A. Bosch, A. Zisserman, and X. Munoz, "Image classification using random forests and ferns," Proc. 11th International Conference on Computer Vision, Rio de Janeiro, Brazil, pp.1–8, 2007.