A Novel Manifold Regularized Online Semi-supervised Learning Algorithm

Shuguang Ding[1], Xuanyang Xi[2], Zhiyong Liu[2,3(✉)], Hong Qiao[2,3], and Bo Zhang[1]

[1] LSEC, Institute of Applied Mathematics, AMSS, Chinese Academy of Sciences, Beijing 100190, China

[2] State Key Lab of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
zhiyong.liu@ia.ac.cn

[3] CAS Centre for Excellence in Brain Science and Intelligence Technology (CEBSIT), Shanghai 200031, China

Abstract. In this paper, we propose a novel manifold regularized online semi-supervised learning (OS^2L) model in an Reproducing Kernel Hilbert Space (RK-HS). The proposed algorithm, named Model-Based Online Manifold Regularization (MOMR), is derived by solving a constrained optimization problem, which is different from the stochastic gradient algorithm used for solving the online version of the primal problem of Laplacian support vector machine (LapSVM). Taking advantage of the convex property of the proposed model, an exact solution can be obtained iteratively by solving its Lagrange dual problem. Furthermore, a buffering strategy is introduced to improve the computational efficiency of the algorithm. Finally, the proposed algorithm is experimentally shown to have a comparable performance to the standard batch manifold regularization algorithm.

Keywords: Manifold regularization · Online semi-supervised learning · Lagrange dual problem

1 Introduction

Cognitive science has drawn a lot of attentions for its significance in understanding human categorization in recent years [5]. In human learning, learners can incrementally learn the classes of various objects from the surrounding environment, where only a few objects are labeled by a knowledgeable source. This scenario can be regarded as online semi-supervised learning, that is, the label of a new arrived sample is unavailable or presented very sporadically in the online process.

In this paper, we focus on the online semi-supervised learning (OS^2L) problems. Several online semi-supervised learning algorithms have been proposed in the past several years. By using a heuristic method to greedily label the unlabeled examples, Babenko et al. [1] and Grabner et al. [9] tried to solve the OS^2L

A. Hirose et al. (Eds.): ICONIP 2016, Part I, LNCS 9947, pp. 597–605, 2016.
DOI: 10.1007/978-3-319-46687-3_66

problems in an online supervised learning framework. Dyer et al. [3] presented a semi-supervised learning (SSL) framework called COMPOSE (COMPacted Object Sample Extraction), where a few labeled samples are given initially, and then a SSL problem is solved based on the currently labeled samples and new unlabeled samples, which are from a drift distribution. To reduce the computational complexity of manifold construction in the online training process, Kveton et al. [11] and Farajtabar et al. [4] proposed the harmonic solution for manifold regularization on an approximate graph. By using online convex programming, Goldberg et al. [6] proposed an online manifold learning framework for SSL in a kernel space with stochastic gradient descent. In addition, they extended their method to online active learning by adding an optional component to select which instances to label [8]. Sun et al. [7,14] exploited the property of Fenchel conjugate of hinge loss and gradient ascend method to solve the dual problem of their online manifold learning model. These algorithms in [6,7,14] are derived by using online gradient methods, implying that these methods can be regarded as solving the off-line semi-supervised learning models by stochastic gradient methods. However, none of these stochastic gradient methods can obtain exact solution because they do not directly solve the constrained optimization problem involved.

Note that an algorithm with an exact solution can obtain better performance. Therefore, to exploit the internal geometry information of the unlabeled data and take advantage of the kernel methods, in this paper we propose a novel online manifold regularization learning model in an Reproducing Kernel Hilbert Space (RKHS). In each iteration of online training, by considering the new arrived sample and the previous samples, an online model based on a constrained optimization problem is presented. Unlike the stochastic gradient method for solving the off-line model, the exact solution of the proposed model can be obtained by exploiting the Lagrange dual problem. In addition, the regularization parameter of the proposed model can be regarded as a forgetting factor, which can be used to control the number of support vectors by considering a buffering strategy in the online leaning process. By such merit, the proposed online predictor experimentally exhibits a high accuracy comparable to batch algorithm LapSVM.

The rest of this paper is organized as follows. Section 2 presents the proposed model and algorithm. Experimental results on several data sets are shown in Sect. 3. Some concluding remarks are given in Sect. 4.

2 Online Manifold Learning

In this section, the proposed model is presented in detail. In Sect. 2.1, a new model is proposed for online manifold regularization learning in an RKHS. In Sect. 2.2, the proposed model is solved by exploiting the property of Lagrange dual problem.

2.1 Online Model Based on Manifold Regularization

Assume that the current learning data for semi-supervised learning are $(x_1, y_1, \delta_1), (x_2, y_2, \delta_2), \ldots, (x_t, y_t, \delta_t)$ where $x_i \in \mathcal{X}$ is a point, $y_i \in \mathcal{Y} = \{-1, 1\}$ is its label and δ_i is a flag to determine whether the label y_i is available (y_i is available if and only if $\delta_i = 1$). At round t, the current predictor is $h_t(x) = \text{sign}(f_t(x))$ and f_0 is set as $f_0 = 0$ in our algorithm. In online semi-supervised learning, when a new sample $(x_{t+1}, y_{t+1}, \delta_{t+1})$ is available, the function f_{t+1} is updated based on the current decision function f_t and the implicit feedback, that is, the manifold structure of the samples.

Suppose that $K(\cdot, \cdot)$ is a chosen Kernel function over the training samples and $\mathcal{H}$ is the corresponding RKHS. Therefore, according to the Representer Theory [13], f_t and f_{t+1} can be written as:

$$f_t(\cdot) = \sum_{i=1}^{t} \alpha_i^t K(x_i, \cdot), f_{t+1}(\cdot) = \sum_{i=1}^{t} \alpha_i^{t+1} K(x_i, \cdot) + \alpha_{t+1}^{t+1} K(x_{t+1}, \cdot). \quad (1)$$

In the online learning process, our aim is to update $\{\alpha_i^{t+1}\}_{i=1}^{t+1}$ from $\{\alpha_i^t\}_{i=1}^t$ based on a proper algorithm. Considering the trade-off between the amount of progress made on each round and the amount of information retained from previous rounds, and compromise the classification error, the manifold constraint and the complexity of f as LapSVM, our online semi-supervised learning model with manifold regularization is presented as:

$$\begin{aligned} \min_{f, \xi_{t+1}} \frac{1}{2}\|f - f_t\|_{\mathcal{H}}^2 + \frac{\lambda_1}{2}\|f\|_{\mathcal{H}}^2 + C\delta_{t+1}\xi_{t+1} + \frac{1}{2}\lambda_2 \sum_{i=1}^{t} (f(x_i) - f(x_{t+1}))^2 w_{it+1} \\ \text{s.t.} \quad y_{t+1} f(x_{t+1}) \geq 1 - \xi_{t+1}, \xi_{t+1} \geq 0 \end{aligned} \quad (2)$$

where $\frac{1}{2}\|f - f_t\|_{\mathcal{H}}^2$ measures the difference between f and the previous f_t, the term $\|f\|_{\mathcal{H}}^2$ controls the complexity of the decision function f, $\sum(f(x_i) - f(x_{t+1}))^2 w_{it+1}$ is the manifold regularizer which depends on the edge weight w_{it+1}, f and x_i, and ξ_{t+1} is the slack variable denoting a possible error for the newly arrived data$(x_{t+1}, y_{t+1}, \delta_{t+1})$ after f is determined, λ_1, λ_2 and C are parameters reflecting the weights compromising complexity, the manifold regularizer and the classification error.

In the objective function of (2), the manifold structure of the samples is reflected in the term $\sum_{i=1}^{t}(f(x_i) - f(x_{t+1}))^2 w_{it+1}$, which can be regarded as an implicit feedback. This regularization term makes the new sample gain a similar decision value to its close sample in the manifold. Therefore, the proposed model can take advantage of the implicit feedback and the kernel methods. The solution of the proposed model is presented in the next section.

2.2 Online Algorithm of the Proposed Model

In this section, we give a detailed solution of the proposed model by exploiting the property of Lagrange dual problem. Assuming that $\delta_{t+1} = 1$ (if $\delta_{t+1} = 0$, the solution of (2) can be obtained by the similar process as below), the Lagrange dual problem of (2) is

$$\begin{aligned} \max_{\gamma_{t+1}} \min_{f,\xi_{t+1}} \quad & L(f, \xi_{t+1}, \gamma_{t+1}, \beta_{t+1}) \\ \text{s.t.} \quad & \gamma_{t+1} \geq 0, \quad \beta_{t+1} \geq 0 \end{aligned} \tag{3}$$

where γ_{t+1} and β_{t+1} are the Lagrange multipliers corresponding to the constraints $y_{t+1} f(x_{t+1}) \geq 1 - \xi_{t+1}$ and $\xi_{t+1} \geq 0$, respectively.

For simplicity, we define D and W as

$$D_{ij} = \begin{cases} w_{ij} & \text{if} \quad 0 < i = j < t+1 \\ \sum_{i=1}^{t} w_{it+1} & \text{if} \quad i = j = t+1 \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

$$W_{ij} = \begin{cases} w_{ij} & \text{if} \quad 0 < i < t+1, j = t+1 \\ w_{ij} & \text{if} \quad i = t+1, 0 < j < t+1 \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

Substituting (1), (4), (5) into (3) and let $L = D - W$, we have

$$\begin{aligned} L(\alpha, \xi_{t+1}, \gamma_{t+1}, \beta_{t+1}) = & \frac{1}{2} \alpha^T (K + \lambda_1 K + \lambda_2 KLK) \alpha \\ & - \gamma_{t+1} (y_{t+1} \alpha^T J - 1 + \xi_{t+1}) + c_0 \\ & - \alpha^T K \tilde{\alpha}^t - \beta_{t+1} \xi_{t+1} + C \xi_{t+1} \end{aligned} \tag{6}$$

where $\alpha = [\alpha_1, \ldots, \alpha_{t+1}]^T$, $\tilde{\alpha}^t = [\alpha_1^t, \ldots, \alpha_t^t, 0]^T$, K is a $(t+1) \times (t+1)$ Gram Matrix with $K_{ij} = K(x_i, x_j)$, $J = Ke$, $e = [0, \ldots, 0, 1]^T$ is a $(t+1)$-dimensional vector and c_0 is a constant.

Note that $L(\alpha, \xi_{t+1}, \gamma_{t+1}, \beta_{t+1})$ attains its minimum with respect to α and ξ_{t+1}, if and only if the following conditions are satisfied:

$$\nabla_\alpha L(\alpha, \xi_{t+1}, \gamma_{t+1}, \beta_{t+1}) = 0, \tag{7}$$

$$\nabla_{\xi_{t+1}} L(\alpha, \xi_{t+1}, \gamma_{t+1}, \beta_{t+1}) = 0. \tag{8}$$

Therefore, we formulate a reduced Lagrangian:

$$\begin{aligned} L^R(\alpha, \gamma_{t+1}) = & \frac{1}{2} \alpha^T (K + \lambda_1 K + \lambda_2 KLK) \alpha + c_0 \\ & - \gamma_{t+1} (y_{t+1} \alpha^T J - 1) - \alpha^T K \tilde{\alpha}^t. \end{aligned} \tag{9}$$

Taking derivative of the reduced Lagrangian with respect to α, we have:

$$\alpha = (K + \lambda_1 K + \lambda_2 KLK)^{-1} (K \tilde{\alpha}^t + J y_{t+1} \gamma_{t+1}). \tag{10}$$

Substituting back in the reduced Lagrangian we get:

$$\begin{aligned} \max_{\gamma_{t+1}} & -\frac{1}{2}(K\tilde{\alpha}^t + Jy_{t+1}\gamma_{t+1})^T A^{-1}(K\tilde{\alpha}^t + Jy_{t+1}\gamma_{t+1}) + \gamma_{t+1} \\ \text{s.t.} & \qquad\qquad 0 \le \gamma_{t+1} \le C, \end{aligned} \tag{11}$$

where $A = K + \lambda_1 K + \lambda_2 KLK$.

Let $\overline{\gamma}_{t+1}$ be the stationary point of the object function of (11).

Therefore

$$\overline{\gamma}_{t+1} = \frac{1 - y_{t+1}J^T A^{-1} K\tilde{\alpha}^t}{J^T A^{-1} J}. \tag{12}$$

Assume that optimal solution of (11) is γ^*_{t+1}. Note that the object function (11) is quadratic, so the optimal solution γ^*_{t+1} in the interval $[0, C]$ is at either 0, C or $\overline{\gamma}_{t+1}$. Hence

$$\gamma^*_{t+1} = \begin{cases} 0, & \text{if} \quad \overline{\gamma}_{t+1} \le 0 \\ C, & \text{if} \quad \overline{\gamma}_{t+1} \ge 0 \\ \overline{\gamma}_{t+1}, & \text{otherwise} \end{cases} \tag{13}$$

Furthermore, if $\delta_{t+1} = 0$, we can obtain the solution of the proposed model by the similar process as above. Thus, the online manifold regularization for classification is presented as

$$\begin{aligned} f_{t+1}(x) &= \sum_{i=1}^{t+1} \alpha_i^{t+1} K(x_{t+1}, x), \\ h_{t+1} &= \text{sign}(f_{t+1}(x)), \end{aligned} \tag{14}$$

where

$$\alpha^{t+1} = A^{-1}(K\tilde{\alpha}^t + \delta_{t+1} y_{t+1} \gamma^*_{t+1} J)$$

The above process of solving the proposed model is denoted by MOMR. In practice, the parameter λ_1 can be regard as a forgetting factor. Suppose λ_2 is very small and $\lambda_1 > 0$. According to (10), we have $\alpha \simeq (1+\lambda_1)^{-1}(\tilde{\alpha}^t + y_{t+1} e \gamma_{t+1})$, that is, $\alpha_i^{t+1} \simeq \alpha_i^t/(1+\lambda_1)$ for $i = 1, \ldots, t$, which means that the absolute value of α_i^t will continually decrease in the online process. Thus, if the absolute value of a coefficient is small in the current decision function, the corresponding sample can be deleted safely from the current support vectors set. Based on this, a buffering strategy is introduced to make the online training feasible in the RKHS. The detailed process of the buffering strategy is presented in Sect. 3.

3 Experiments

To verify the effectiveness, we compare the proposed algorithm with two online manifold regularization algorithms and a batch algorithm on two data sets. In Sect. 3.1, the experimental setups are introduced in detail. In Sect. 3.2, several experiments are processed and the results are summarized and analyzed in detail.

3.1 Experimental Setup

Two data sets are used in our experiments. The first data set is the MNIST [12]. We focus on the binary classification task of separating '6' from '8' (MNIST6VS8) in our experiment. The sizes of the training set and test set are 11769 and 1932 respectively. The second data set is the FACEMIT [10] which contains 361-dimensional images of faces and non-faces. A balanced subset from FACEMIT (size 5000) is sampled and divided into two sets: the training set and the test set with a proportion 1:1 for our experiment. Similar to the experimental settings in [6,7], the labeled rate of the training samples is set to be 2 % in all the experiments.

In our experiments, we focus on online manifold regularization algorithms derived from the dual problem. Therefore, We compare the performance of our algorithm MOMR with an online manifold regularization algorithm based on Example-Associate Update (denoted by OMR-EA), an online manifold regularization algorithm based on Overall Update (denoted by OMR-Overall) [7] and a batch manifold regularization algorithm LapSVM [2].

To reduce the storage for online learning in an RKHS, we use a buffering strategy for all the online algorithms: Let the buffer size be B. If the buffer is full, the sample with the smallest absolute coefficient in the buffer is replaced by the new arrived sample. We evaluate the three online algorithms separately with different buffer sizes (B $\in \{50, 200\}$) in our experiments.

In all the experiments, the RBF kernel $k(x_i, x_j) = \exp(-\|x_i - x_j\|^2/(2\sigma_K^2))$ is used for classification and the edge weights are Gaussian weights $k(x_i, x_j) = \exp\,(-\|x_i - x_j\|^2/(2\sigma_W^2))$ which define a fully connected graph. The parameter values $\sigma_K, \sigma_W, \lambda_1$ and λ_2 are selected by using 5-fold cross validation on the first 500 samples of the training data, where $\sigma_K, \sigma_W \in \{2^{-3}, 2^{-2}, 2^{-1}, 2^0,\ 2^1, 2^2, 2^3\}$ and $\lambda_1, \lambda_2 \in \{10^{-5}, 10^{-4}, 10^{-3},\ 10^{-2}, 10^{-1}, 10^0, 10^1, 10^2\}$. In addition, as suggested in [7], the step sizes of the OMR-EA and OMR-Overall are set to be a small value 0.01. The value of parameter C is set to be 1 for the proposed algorithm MOMR. The computational efficiencies of all the algorithms are evaluated in terms of their CPU running time (in seconds). All the experiments are implemented in Matlab over a desktop PC with Inter(R) Core(TM) 3.2 GHz CPU, 4G RAM and Windows 7 operating system.

3.2 Online Processing and Performance Evaluation

In this subsection, we give out a detailed process of the experiments and evaluate the performance of the proposed algorithm for online manifold regularization learning.

All the three online algorithms are performed in the same way which are divided into two steps: (1) Online processing. Train a classifier with a new arrived sample; (2) Test. Test the final model on a test set. In each learning round, the batch algorithm LapSVM is trained with all the visible samples. We repeat all the experiments 10 times (each with an independent random permutation of the training samples) and the results presented below are all average over 10 trials.

Table 1. The accuracy of different algorithms on the data set MNIST6VS8 and FACEMIT with different buffer sizes. The best classification results are marked in boldface.

Date set	B	MOMR	OMR-EA	OMR-Overall	LapSVM
MNIST6VS8	50	98.012 ± 0.442	96.491 ± 1.775	97.495 ± 0.714	98.861 ± 0
MNIST6VS8	200	$\mathbf{99.048 \pm 0.078}$	98.954 ± 0.177	97.981 ± 0.543	98.861 ± 0
FACEMIT	50	$\mathbf{78.024 \pm 3.411}$	77.992 ± 3.390	78.000 ± 3.478	77.600 ± 0
FACEMIT	200	$\mathbf{78.552 \pm 3.360}$	77.948 ± 3.126	77.920 ± 3.237	77.600 ± 0

The test accuracies on the two data sets are summarized in Table 1. From the results, the test accuracy of MOMR is comparable with the off-line algorithm LapSVM on the two data sets and higher than those of the two online algorithms OMR-EA and OMR-Overall. These are reasonable since that: (a) in our algorithm, the exact solution is obtained from the proposed model; (b) in OMR-EA and OMR-Overall, the approximate solutions of their models are derived by online gradient method.

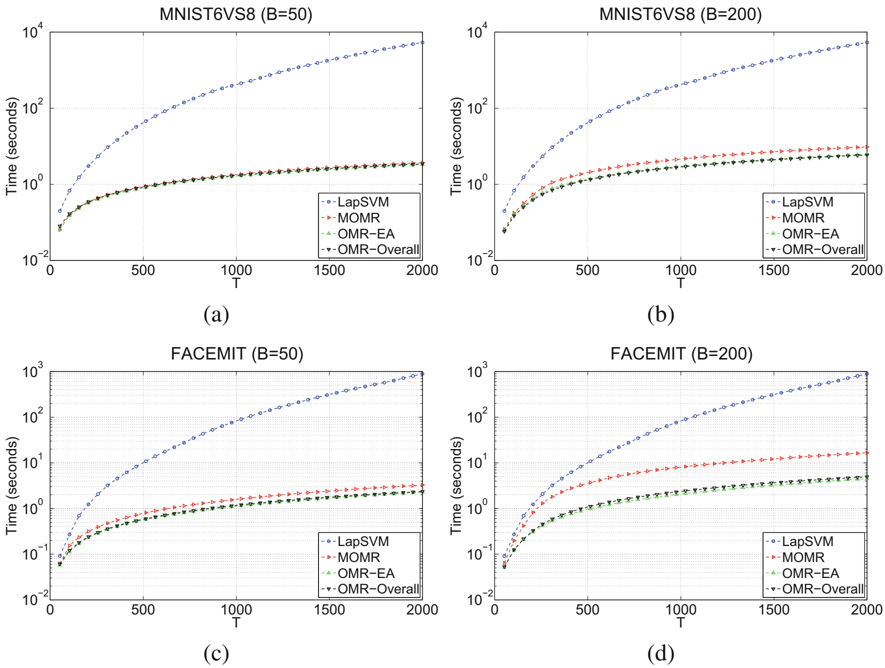

Fig. 1. Cumulative running time of online updating the classifiers with different buffer sizes on the data set MNIST6VS8 and FACEMIT.

The online updating time of the four algorithms are presented in Fig. 1. With respect to the running time, we can see that MOMR is comparable to the online algorithms OMR-EA and OMR-Overall when the buffer size is small and much faster than the off-line algorithm LapSVM. These can be explained by: (a) each sample is only trained once by the online algorithms; (b) a buffering strategy is used to reduce the repeated training process.

Considering above two results, it can be inferred that the proposed algorithm is in the first grade among the four algorithms both on the test accuracy aspect and on the running time aspect.

Additionally, in practice, the buffer size can be used to trade-off the accuracy and the time cost of online classifiers. The appropriate buffer size can be derived by using cross validation on the first N arrived samples, where N is a predefined number.

4 Conclusion

In this paper, the proposed model offers a new method to solve the OS^2L problem. Experiment results verify the effectiveness of the proposed algorithm. In addition, the proposed method enriches the research fields of cognitive computation and LapSVM.

Acknowledgment. This work is partly supported by NFSC grants 61375005 and MOST grants 2015BAK-35B01.

References

1. Babenko, B., Yang, M.H., Belongie, S.: Visual tracking with online multiple instance learning. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 983–990 (2009)
2. Belkin, M., Niyogi, P., Sindhwani, V.: Manifold regularization: a geometric framework for learning from labeled and unlabeled examples. J. Mach. Learn. Res. **7**, 2399–2434 (2006)
3. Dyer, K.B., Capo, R., Polikar, R.: Compose: a semisupervised learning framework for initially labeled nonstationary streaming data. IEEE Trans. Neural Netw. Learn. Syst. **25**(1), 12–26 (2014)
4. Farajtabar, M., Shaban, A., Rabiee, H.R., Rohban, M.H.: Manifold coarse graining for online semi-supervised learning. In: Gunopulos, D., Hofmann, T., Malerba, D., Vazirgiannis, M. (eds.) ECML PKDD 2011. LNCS (LNAI), vol. 6911, pp. 391–406. Springer, Heidelberg (2011). doi:10.1007/978-3-642-23780-5_35
5. Gibson, B.R., Rogers, T.T., Zhu, X.: Human semi-supervised learning. Top. Cogn. Sci. **5**(1), 132–172 (2013)
6. Goldberg, A.B., Li, M., Zhu, X.: Online manifold regularization: a new learning setting and empirical study. In: Daelemans, W., Goethals, B., Morik, K. (eds.) ECML PKDD 2008. LNCS (LNAI), vol. 5211, pp. 393–407. Springer, Heidelberg (2008). doi:10.1007/978-3-540-87479-9_44
7. Sun, B.L., Li, G.H., Jia, L., Zhang, H.: Online manifold regularization by dual ascending procedure. Math. Probl. Eng. **2013** (2013). doi:10.1155/2013/838439

8. Goldberg, A.B., Zhu, X.J., Furger, A., Xu, J.M.: Oasis: online active semi-supervised learning. In: Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence, pp. 1–6 (2011)
9. Grabner, H., Leistner, C., Bischof, H.: Semi-supervised on-line boosting for robust tracking. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008. LNCS, vol. 5302, pp. 234–247. Springer, Heidelberg (2008). doi:10.1007/978-3-540-88682-2_19
10. Heisele, B., Poggio, T., Pontil, M.: Face detection in still gray images. AI Memo 1697, Center for Biological and Computational Learning, MIT, Cambridge, MA (2000)
11. Kveton, B., Philipose, M., Valko, M., Huang, L.: Online semi-supervised perception: real-time learning without explicit feedback. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 15–21 (2010)
12. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proc. IEEE **86**(11), 2278–2324 (1998)
13. Schölkopf, B., Herbrich, R., Smola, A.J.: A generalized representer theorem. In: Helmbold, D., Williamson, B. (eds.) COLT 2001. LNCS (LNAI), vol. 2111, pp. 416–426. Springer, Heidelberg (2001). doi:10.1007/3-540-44581-1_27
14. Sun, B.L., Li, G.H., Jia, L., Huang, K.H.: Online coregularization for multiview semisupervised learning. Sci. World J. **2013** (2013). doi:10.1155/2013/398146