Journal of Applied Sciences

ISSN 1812-5654

science alert

ANSInet
an open access publisher
http://ansinet.com

Journal of Applied Sciences 13 (10): 1728-1733, 2013
ISSN 1812-5654 / DOI: 10.3923/jas.2013.1728.1733

An Effective Active Semi-supervised Learning Method Based on Manifold Regularization

[1,2]Xiukuan Zhao, [1,2]Baiqi Ning and [3]Gangbing Song
[1]Key Laboratory of Ionospheric Environment, Institute of Geology and Geophysics,
Chinese Academy of Sciences, Beijing, 100029, China
[2]Beijing National Observatory of Space Environment, Institute of Geology and Geophysics,
Chinese Academy of Sciences, Beijing, 100029, China
[3]Department of Mechanical Engineering, University of Houston, Houston, TX 77204, United State of America

Abstract: Conventional artificial intelligent methods such as neural network and SVM use only labeled data (feature/label pairs) for training. Labeled instances are often difficult, expensive, or time consuming to obtain. To use a large amount of unlabeled data together with labeled data to build better models, we proposed an active semi-supervised learning method based on the pool query active learning and manifold regularization semi-supervised method. In this paper, the effectiveness of the method was verified by its application to a synthetic data set and three real world classification problems. The experimental results showed that employing our active semi-supervised learning method can significantly reduce the need for labeled training instances.

Keywords: Semi-supervised learning, active learning, LapSVM, A-LapSVM

INTRODUCTION

In many supervised learning tasks, labeled instances are often difficult, expensive, or time consuming to obtain. Meanwhile, unlabeled data may be relatively easy to collect however difficult to use. Thus, finding ways to minimize the number of labeled instances is beneficial. Usually, the training set is chosen to be a random sampling of instances. However, in many cases active learning can be employed. Here, the learner can actively choose the training data. The intention here is for active learning to allow the learner this extra flexibility which would reduce the learner's need for large quantities of labeled data. In recent years, many methods that can be broadly divided into two groups, semi-supervised and active learning, have been proposed to solve such problems.

Semi-supervised learning algorithms use large amounts of unlabeled data, together with the labeled data, to build better classifiers. They are mainly based on three paradigms: density based methods (Chapelle *et al*., 2008), graph-based algorithms (Belkin *et al*., 2006; Weston *et al*., 2012) and boosting techniques (Saffari *et al*., 2009). There are many applications using the semi-supervised learning method, such as human action recognition (Zhao *et al*., 2013) and diabetes diseases prediction (Wu *et al*., 2009). A survey on semi-supervised learning is presented by Zhu (2008).

The objective of active learning is to learn a function that accurately predicts the labels of new examples while requesting as few labels as possible. It has been studied for many real-world problem domains in machine learning, such as text classification (Tong and Koller, 2001), image classification and retrieval (Shen *et al*., 2011; Zhang and Chen, 2002), video indexing (Zha *et al*., 2012) and cancer diagnosis (Liu, 2004). Pool-based active learning appears to be one of the more common types among application papers. Settles (2010) presents a comprehensive survey about the literature of active learning.

Both semi-supervised learning and active learning take advantage of the unlabeled data, thus, it is quite natural to combine them to form a more effective method. Zhu *et al*. (2003) proposed an approach to couple active learning with semi-supervised learning using Gaussian fields and harmonic functions. Yu *et al*. (2010) proposed a unified Global Entropy Reduction Maximization (GERM) framework for active learning and semi-supervised learning for speech recognition. We proposed an active semi-supervised learning method (Zhao *et al*., 2011) and call it A-LapSVM. It was proposed based on the pool query active learning and manifold regularization semi-supervised method which is a graph based method.

In this study, the method is extended to be used in several domains. The procedure of the active semi-supervised method is introduced in section 2. In section 3 we present experimental results for one

Corresponding Author: Xiukuan Zhao, Key Laboratory of Ionospheric Environment, Institute of Geology and Geophysics, Chinese Academy of Sciences, Beijing, China

synthetic dataset and three real world classification problems. Finally, we offer our conclusions in section 4.

ACTIVE SEMI-SUPERVISED LEARNING

Suppose we have a training set of labeled and unlabeled samples and the number of labeled samples is too few for supervised methods to build a classifier with a reasonable level of performance. In this case, the unlabeled data can help to build a classifier with greater accuracy. The goal of active learning with a semi-supervised classifier is to query the unlabeled samples that reveals the most information while taking into account the information already provided by the pool of unlabeled samples.

Semi-supervised learning using manifold regularization: Graph-based semi-supervised methods define a graph where the nodes are labeled and the unlabeled examples in the dataset and edges (may be weighted) reflect the similarity of the examples. These methods usually assume label smoothness over the graph. The manifold regularization framework is a kind of graph based method. It employs two regularization terms (Belkin *et al*., 2006):

$$\min \frac{1}{l}\sum_{i=1}^{l} V(x_i, y_i, f) + \gamma_A \|f\|_K^2 + \gamma_I \|f\|_I^2$$

where, V is an arbitrary loss function, K is a Mercer kernel. $\|f\|_K$ is an appropriate penalty term that should reflect the intrinsic structure. $\|f\|_K$ is a smoothness penalty corresponding to the probability distribution. γ_A controls the complexity of the function in the ambient space while γ_I controls the complexity of the function in the intrinsic geometry.

By using the manifold regularization framework, the SVM can be extended to a Laplacian SVM (LapSVM) by solving the following problem (Belkin *et al*., 2006):

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

where, $f = [f(x_1),\cdots f(x_{l+u})]^T$ and L is the graph Laplacian given by L = D-W. The diagonal matrix D is given by:

$$D_{ii} = \sum_{j=1}^{l+u} W_{ij}$$

and W_{ij} is edge weight in the data adjacency graph.

The standard SVM equation is extended as:

$$\min_{\alpha\in\mathbb{R}^{l+u}, \xi\in\mathbb{R}^l} \frac{1}{l}\sum_{i=1}^{l}\xi_i + \gamma_A \alpha^T K\alpha + \frac{\gamma_I}{(u+1)^2}\alpha^T KLK\alpha$$

$$\text{s.t. } y_i\left(\sum_{j=1}^{l+u}\alpha_j K(x_i, x_j) + b\right) \geq 1-\xi_i, i=1,\ldots,l, \xi_i \geq 0, i=1,\ldots,l$$

The dual form of Eq. 3 is formed as:

$$L_D = \max_{\beta\in\mathbb{R}^l} \sum_{i=1}^{l}\beta_i - \frac{1}{2}\beta^T Q\beta$$

$$\text{s.t. } 0 \leq \beta_i \leq \frac{1}{l}, i=1,\ldots,l \text{ and } \sum_{i=1}^{l}\beta_i y_i = 0$$

Where:

$$Q = YJK\left(2\gamma_A I + 2\frac{\gamma_I}{(u+1)^2}LK\right)^{-1} J^T Y$$

The LapSVM is implemented by using a standard SVM solver with the quadratic form induced by Eq. 4.

The decision function is:

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

Experimental evidence and theoretical analysis suggest that the LapSVM algorithm is able to use unlabeled data effectively and obtain more accurate classifier (Belkin *et al*., 2006).

Active semi-supervised learning: Given an unlabeled pool U and a labeled dataset L, an active semi-supervised learner l has three components: (h, q, L∪U). The first component h is a semi-supervised classifier, trained on the current set of the labeled data L and the unlabeled samples U. The second component q is the query function that decides which sample in U to query next. The active semi-supervised learner returns a classifier h after each query. The main difference between an active learner and a passive one is the query component q ($q = \arg\min_{x\in U}|h(x)|$). This brings us to the issue of choosing the next unlabeled sample to query. The difference between active semi-supervised learning and active supervised learning is whether or not using the samples in unlabeled pool U to train h. The active semi-supervised learning algorithm is outlined in Table 1.

The diagram of the active semi-supervised learning method is shown in Fig. 1. The data source contains both labeled data and unlabeled data are used in semi-supervised learning. The active learning is processed based on the output of semi-supervised learning; it will select the informative data to let the oracle (e.g., a human expert) to give the label. The new labeled data is then used in the semi-supervised learning. The process iterates until the condition is satisfied and the output is then the trained model. We call this active learning procedure with LapSVM solution Active LapSVM (A-LapSVM).

Table 1: Active semi-supervised learning algorithm

Given: labeled dataset L_0: $(x_1, y_1),\cdots,(x_l, y_l)$ where $x_i \in X$, $y_i \in Y$, unlabeled dataset U_0: $x_{l+1},\cdots,x_{l+u}$ where $x_i \in X$

Repeat

- Train semi-supervised classifier h_j using the current set of labeled dataset L_j and unlabeled samples U_j
- Choose the unlabeled sample x_j:

$$\arg\min_{x \in U_j} \left| w_j \cdot \Phi(x) \right|$$

$\Phi(x)$ is the feature vector and w_j represents the hyperplane of current classifier

- Label the example x_j with label y_j
- Update L_j+2: L_j add (x_j, y_j), U_{j+1}:U_j subtract x_j
- Until stopping criteria is met

Output: active semi-supervised learning classifier h_{final}

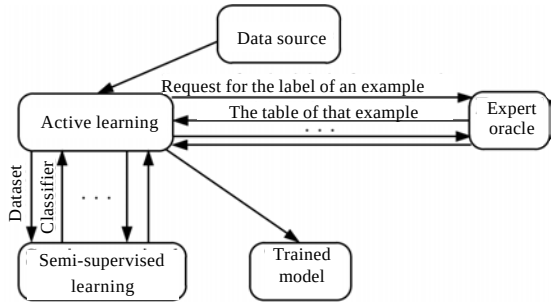

Fig. 1: Diagram of active semi-supervised learning

EXPERIMENTS

We made experiments on a synthetic data set and three real world classification problems arising in credit approval, diabetes categorization and heart disease categorization. All these experiments are binary classification problems. Comparisons were made with two active methods (A-LapSVM, A-SVM (Tong and Koller, 2001) and two random query methods (R-LapSVM, R-SVM). The two random query methods should randomly select one unlabeled data in U to request for labeling.

Two moons dataset: We performed the A-LapSVM on the two moons dataset (Fig. 2). The data set contained 200 examples with only 1 labeled example for each class at the very beginning. We chose the RBF kernel as the kernel function and the kernel parameter was 0.35. The parameters of LapSVM were set to $\gamma_A = 0.2$, $\gamma_I = 0.5$.

Figure 2 demonstrates how A-LapSVM successfully finds the most informative example to query next. In Fig. 2, the blue diamond denotes a positive type, the magenta round denotes a negative type. The black points represent unlabeled examples and the red cross represents the active selected example. Also shown are the decision surfaces of the A-LapSVM for each step. We only need

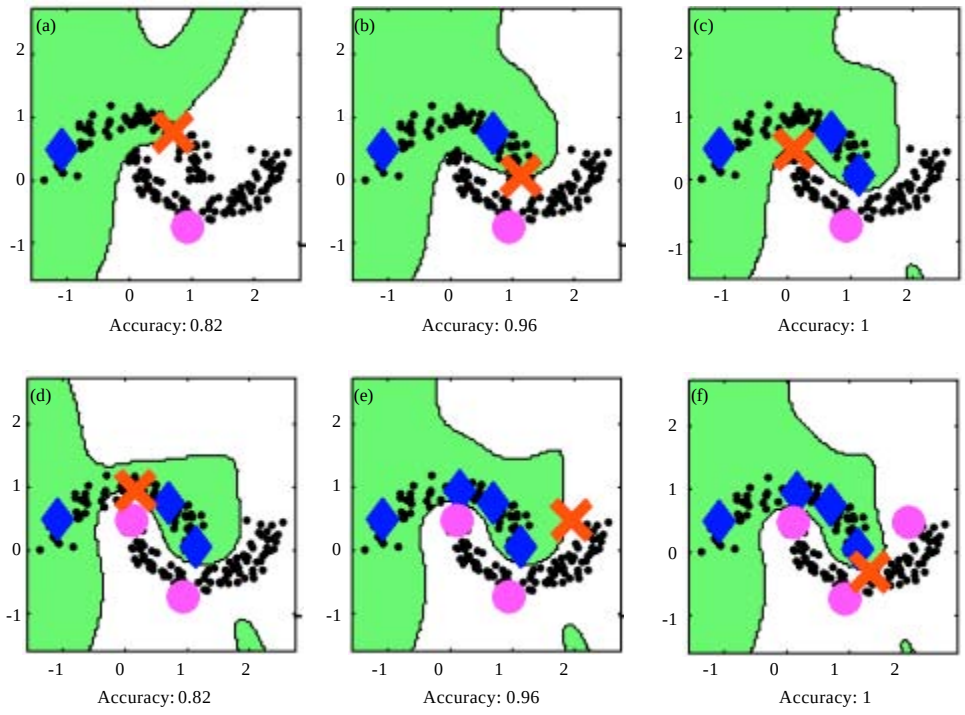

Fig. 2(a-f): Procedure of A-LapSVM with two moons data set

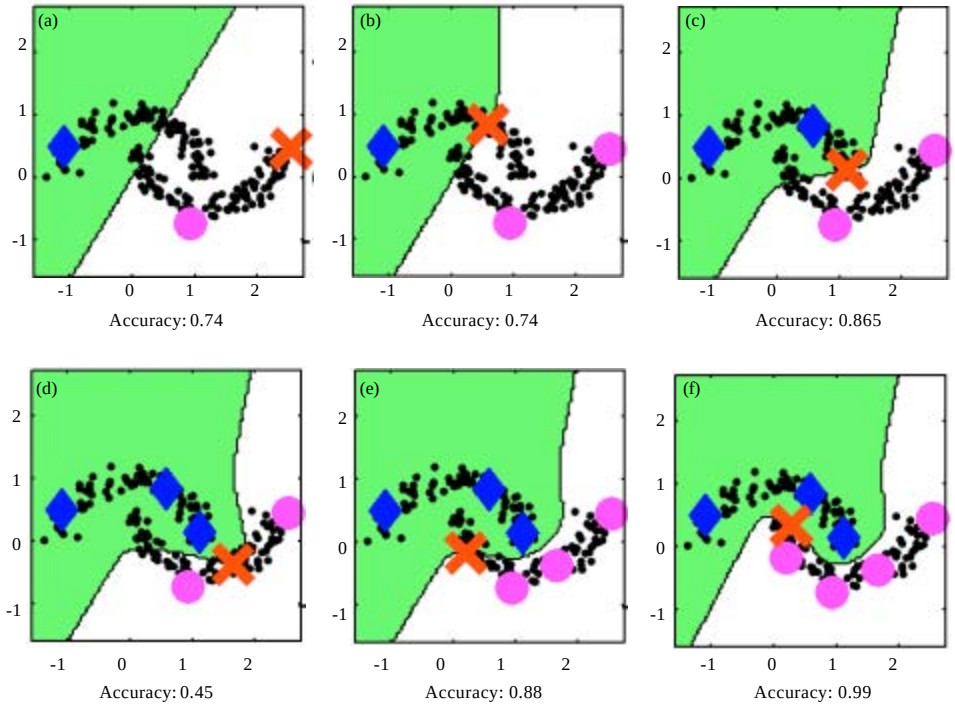

Fig. 3(a-f): Procedure of Active SVM with two moons data set

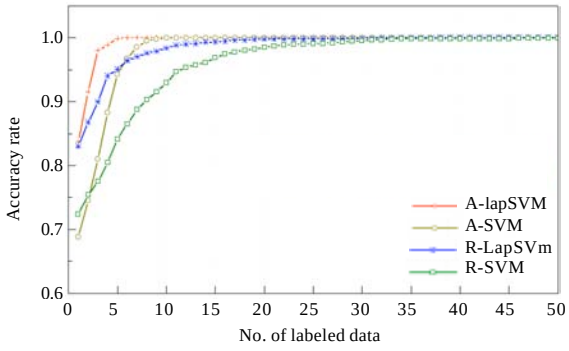

Fig. 4: Accuracy rate comparison diagram of two moons experiment

to label quit few examples, not all of the 200 examples. In the first step, with only one labeled example for each class, the classification accuracy is 0.82. The accuracy in step 2 is 0.96 and in step 3 the data is completely separated. We will achieve the optimal solution after three steps. The A-LapSVM decision boundary seems to be intuitively satisfied from step 3 to 6. In reality, this will reduce the labor cost of labeling to a great extent.

In order to contrast with the A-SVM method, we also plot the procedure of A-SVM with two moons data set in Fig. 3. In the first step, the classification accuracy is 0.74. The accuracy in step 2 is 0.74 and it is 0.865 in step 3. The A-SVM method needs six steps to approximate completely separate all two moons data hwile the A-LapSVM method only needs three steps.

To compare the efficiency of the four different classifier methods, we took 100 cycles to achieve the average accuracy. During each cycle, we randomly chose each class with one labeled data. In Fig. 4, the x-axis represents the number of labeled data sets and y-axis represents the accuracy rate which is the mean value of 100 cycles. Each point indicates the classification accuracy rate corresponding to different number of labeled data. From Fig. 4, we can see the A-LapSVM method achieves the highest accuracy rate using only 5 labeled data hwile A-SVM uses 10 labeled data, R-LapSVM uses 30 labeled data and R-SVM uses almost 50 labeled data. Thus, the A-LapSVM method is the most efficiency method to select informative data to label.

Australian credit approval dataset: This experiment was performed on the Australian Credit Approval dataset (available from the UCI machine learning repository (Bache and Lichman, 2013). The dataset concerns credit card applications. It is interesting because there is a good mix of attributes-continuous, nominal with small numbers

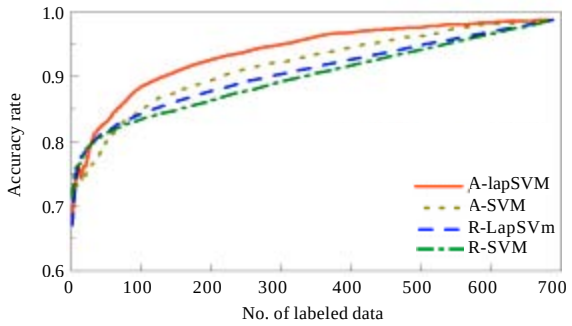

Fig. 5: Accuracy rate comparison diagram of Australian credit approval experiment

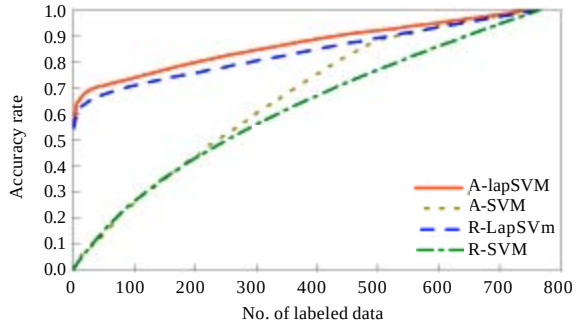

Fig. 6: Accuracy rate comparison diagram of Pima Indians diabetes experiment

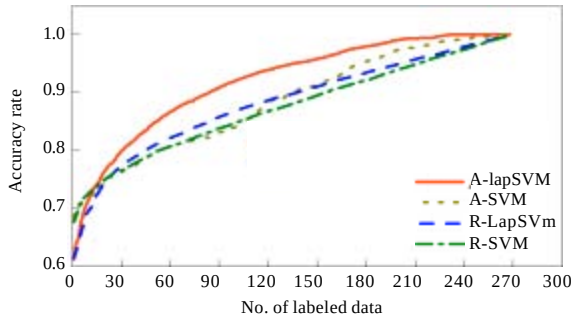

Fig. 7: Accuracy rate comparison diagram of heart disease experiment

of values and nominal with larger numbers of values. The number of instances is 690 and each instance has 14 attributes.

From the very beginning, we supposed that each class had only one labeled example. Then we performed the four query methods (A-LapSVM, A-SVM, R-LapSVM and R-SVM) with this dataset. After each query step, we calculated the accuracy of the current classifier on the whole testing data. We also took 100 cycles to achieve the average accuracy. In Fig. 5, we compare the accuracy rate of the four different methods. In this experiment, the A-LapSVM achieves the highest accuracy rate faster than others. It outperforms other methods during the whole process. For example, when the number of labeled data is 200, the accuracy of A-LapSVM is 0.926 while that of A-SVM is 0.894, R-LapSVM is 0.878 and R-SVM is 0.867. The most significant improvements of our method occur in the midrange, seemingly when the number of labeled points is between 100 and 400.

Pima indians diabetes dataset: The Pima Indians Diabetes dataset contains information on various medical measurements on 768 individuals and an indicator of whether they later developed diabetes. Several constraints are placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage. The dataset has 8 attributes. It is available from the UCI machine learning repository.

We followed the same procedure as with the Australian Credit Approval dataset above. First, one labeled data of each class was random seleted to form labeled dataset L_0. The rest of the dataset was considered to be U_0. After each query step, we calculated the accuracy of the current classifier on the whole testing data. We averaged the results over 100 independent processes. The results are shown in Fig. 6.

We observe that the results are very similar to those for the Australian Credit Approval dataset. The performance of our algorithm A-LapSVM is better than others. For example, the accuracy rate of A-LapSVM is 0.845 when the number of labeled data is 300 hwile that of R-LapSVM is 0.804, A-SVM is 0.602 and R-SVM is 0.560. We also see that the unlabeled points significantly improve the classification error when there are very few labeled points.

Heart disease dataset: This dataset is a heart disease database. It has 270 instances and contains 13 attributes. Each instance can be classed to absence or presence of heart disease. It is also available from the UCI machine learning repository.

We followed the same procedure as before, the results are shown in Fig. 7. We can find that our A-LapSVM method is better than others. For example, the accuruacy rate of A-LapSVM is 0.980 when the number of labeled data is 180 hwile that of A-SVM is 0.954, R-LapSVM is 0.934 and R-SVM is 0.921.

CONCLUSION

We developed an effective method A-LapSVM based on the active learning and semi-supervised method with manifold regularization. The proposed method uses large amounts of unlabeled data, together with the labeled

data, to build better classifiers. The effectiveness of the A-LapSVM method was verified by its application to a synthetic data set and three real world classification problems. The following conclusions can be made by these experiments.

- The active query procedure is useful to select the informative data to label. The A-LapSVM and A-SVM methods outperform R-LapSVM and R-SVM in most stages. For example, when the number of labeled data is 200 in Australian Credit Approval experiment, the accuracy of A-LapSVM is 0.926 and A-SVM is 0.894 hwile that of R-LapSVM is 0.878 and R-SVM is 0.867
- The semi-supervised method can use the unlabeled data to form more accuracy classifier model especially when the labeled data is small. In the Pima Indians Diabetes experiment, the accuruacy rate of A-LapSVM is 0.845 when the number of labeled data is 300 and R-LapSVM is 0.804 hwile that of A-SVM is 0.602 and R-SVM is 0.560
- Compared to the state-of-the-art A-SVM, R-LapSVM and R-SVM, the proposed A-LapSVM achieves better performance especially when there are only a few labeled data; it is an efficient learning technique designed to reduce the labor cost of labeling

ACKNOWLEDGMENT

The reported research was supported by the National Natural Science Foundation of China (No. 41104106, 41074113, 41204113).

REFERENCES

Bache, K. and M. Lichman, 2013. UCI machine learning repository. University of California, School of Information and Computer Science, Irvine, CA.

Belkin, M., P. Niyogi and V. Sindhwani, 2006. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. J. Mach. Learn. Res., 12: 2399-2434.

Chapelle, O., V. Sindhwani and S.S. Keerthi, 2008. Optimization techniques for semi-supervised support vector machines. J. Mach. Learn. Res., 9: 203-233.

Liu, Y., 2004. Active learning with support vector machine applied to gene expression data for cancer classification. J. Chem. Inform. Comput. Sci., 44: 1936-1941.

Saffari, A., C. Leistner and H. Bischof, 2009. Regularized multi-class semi-supervised boosting. Proceedings of IEEE Society Conference on Computer Vision and Pattern Recognition, June 20-25, 2009, Miami Beach, FL., USA.

Settles, B., 2010. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin-Madison, pp: 1-67.

Shen, J., B. Ju, T. Jiang, J. Ren, M. Zheng, C. Yao and L. Li, 2011. Column subset selection for active learning in image classification. Neurocomputing, 74: 3785-3792.

Tong, S. and D. Koller, 2001. Support vector machine active learning with applications to text classification. J. Mach. Learn. Res., 2: 45-66.

Weston, J., F. Ratle, H. Mobahi and R. Collobert, 2012. Deep Learning via Semi-Supervised Embedding. In: Neural Networks: Tricks of the Tradem, Montavon, G., G. Orr and K.R. Muller (Eds.). Springer, USA., pp: 639-655.

Wu, J., Y.B. Diao, M.L. Li, Y.P. Fang and D.C. Ma, 2009. A semi-supervised learning based method: Laplacian support vector machine used in diabetes disease diagnosis. Interdiscip. Sci. Comput. Life Sci., 1: 151-155.

Yu, D., B. Varadarajan, L. Deng and A. Acero, 2010. Active learning and semi-supervised learning for speech recognition: A unified framework using the global entropy reduction maximization criterion. Comput. Speech Language, 24: 433-444.

Zha, Z.J., M. Wang, Y.T. Zheng, Y. Yang, R. Hong and T.S. Chua, 2012. Interactive video indexing with statistical active learning. IEEE Trans. Multimedia, 14: 17-27.

Zhang, C. and T. Chen, 2002. An active learning framework for content-based information retrieval. IEEE Trans. Multimedia, 4: 260-268.

Zhao, X., M. Li, J. Xu and G. Song, 2011. An effective procedure exploiting unlabeled data to build monitoring system. Expert Syst. Appl., 38: 10199-10204.

Zhao, X., X. Li, C. Pang and S. Wang, 2013. Human action recognition based on semi-supervised discriminant analysis with global constraint. Neurocomputing, 105: 45-50.

Zhu, X., J. Lafferty and Z. Ghahramani, 2003. Combining active learning and semi-supervised learning using gaussian fields and harmonic functions. Proceedings of the ICML 2003 Workshop on The Continuum from Labeled to Unlabeled Data in Machine Learning and Data Mining, August 2003, Washington, DC., pp: 58-65.

Zhu, X.J., 2008. Semi-supervised learning literature survey. University of Wisconsin-Madison, Wisconsin, Technology Report and Computer Sciences, TR1530.