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

1, 2Xiukuan Zhao, 1, 2Baiqi Ning and 3Gangbing Song
1Key Laboratory of Ionospheric Environment, Institute of Geology and Geophysics,
Chinese Academy of Sciences, Beijing, 100029, China
2Beijing National Observatory of Space Environment, Institute of Geology and Geophysics,
Chinese Academy of Sciences, Beijing, 100029, China
3Department 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 disease 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
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):

\[
\min \frac{1}{n} \sum_{i=1}^{n} V(x_i, y_i, t) + \gamma_t \frac{\| f \|^2}{2} + \gamma_f \frac{\| t \|^2}{2}
\]

where, \( V \) is an arbitrary loss function, \( K \) is a Mercer kernel. \( \| f \| \) is an appropriate penalty term that should reflect the intrinsic structure. \( \| t \| \) is a smoothness penalty corresponding to the probability distribution. \( \gamma_t \) controls the complexity of the function in the ambient space while \( \gamma_f \) 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):

\[
\min \frac{1}{n} \sum_{i=1}^{n} (1 - y_i f(x_i)) + \gamma_t \frac{\| f \|^2}{2} + \gamma_f \frac{\| t \|^2}{2} + \frac{\gamma_f}{2} \sum_{i=1}^{n} f^2(x_i)
\]

where, \( f = [f(x_1), \ldots, f(x_n)]^T \) and \( L \) is the graph Laplacian given by \( L = D - W \). The diagonal matrix \( D \) is given by:

\[
D_{ii} = \sum_{j} W_{ij}
\]

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

The standard SVM equation is extended as:

\[
\min \sum_{a \in \mathcal{A}} \alpha_a K(x_a, x_a) + \gamma \frac{1}{n} \sum_{i=1}^{n} y_i \alpha_i - \frac{1}{2} \alpha^T K L K \alpha
\]

s.t. \( \sum_{a \in \mathcal{A}} \alpha_a = ba + b \)

The dual form of Eq. 3 is formed as:

\[
L = \max_{\beta} \sum_{i=1}^{n} y_i \beta^T \xi_i \quad \text{s.t. } 0 \leq \beta_i \leq \frac{1}{n} \quad i = 1, \ldots, n \quad \text{and } \sum_{i=1}^{n} \beta_i = \beta
\]

Where:

\[
Q = \text{YK} \left( 2 \gamma + 2 \frac{\gamma_f}{(u + 1)^2} \right)^{\gamma_f} Y
\]

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

The decision function is:

\[
f(x) = \sum_{a \in \mathcal{A}} \alpha_a K(x_a, x)
\]

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

<table>
<thead>
<tr>
<th>Given: labeled dataset $L$, $(x_i, y_i)$, $(x_0, y_0)$ where $x_i \neq x$, $y_i \neq Y$, unlabeled dataset $U$, $x_{n+1}$, $x_{n+2}$, select $x_j$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Repeat</td>
</tr>
<tr>
<td>• Train semi-supervised classifier $h_j$ using the current set of labeled dataset $L_j$ and unlabeled samples $U_j$</td>
</tr>
<tr>
<td>• Choose the unlabeled sample $x_j$</td>
</tr>
<tr>
<td>$\arg \min_{x_j} q_j(\Phi(x_j))$</td>
</tr>
<tr>
<td>$\Phi(x_j)$ is the feature vector and $w_j$ represents the hyperplane of current classifier</td>
</tr>
<tr>
<td>• Label the example $x_j$ with label $y_j$</td>
</tr>
<tr>
<td>• Update $L_j+1$: $L_j$ add $(x_j, y_j)$, $U_j$ subtract $x_j$</td>
</tr>
<tr>
<td>• Until stopping criteria is met</td>
</tr>
<tr>
<td>Output: active semi-supervised learning classifier $h_{n+1}$</td>
</tr>
</tbody>
</table>

**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 $\gamma = 0.2$, $\gamma_i = 0.5$.

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. 1:** Diagram of active semi-supervised learning

**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 while 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 while 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
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 selected to form labeled dataset $L_p$. The rest of the dataset was considered to be $U_p$. 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 while 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 accuracy rate of A-LapSVM is 0.980 when the number of labeled data is 180 while 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 while 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 accurate classifier model especially when the labeled data is small. In the Pima Indians Diabetes experiment, the accuracy rate of A-LapSVM is 0.845 when the number of labeled data is 300 and R-LapSVM is 0.804 while 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


