An AdaBoost algorithm for multiclass semi-supervised learning

Tanha, J.; van Someren, M.; Afsarmanesh, H.

Published in:
12th IEEE International Conference on Data Mining: ICDM 2012, 10-13 December 2012, Brussels, Belgium: proceedings

DOI:
10.1109/ICDM.2012.119

Citation for published version (APA):
Abstract—We present an algorithm for multiclass Semi-Supervised Learning which is learning from a limited amount of labeled data and plenty of unlabeled data. Existing semi-supervised algorithms use approaches such as one-versus-all to convert the multiclass problem to several binary classification problems which is not optimal. We propose a multiclass semi-supervised boosting algorithm that solves multiclass classification problems directly. The algorithm is based on a novel multiclass loss function consisting of the margin cost on labeled data and two regularization terms on labeled and unlabeled data. Experimental results on a number of UCI datasets show that the proposed algorithm performs better than the state-of-the-art boosting algorithms for multiclass semi-supervised learning.

Keywords—Semi-Supervised Learning; boosting; multiclass classification

I. INTRODUCTION

In many applications of Machine Learning, for example in classification of images, texts, and multimedia, it is difficult to obtain values of the target variable, while unlabeled examples are plentiful. Semi-supervised learning algorithms use not only the labeled data but also the unlabeled data to build a classifier. One approach to semi-supervised learning is to extend a boosting algorithm, which is one of the most successful meta-classifiers for supervised learning [4]. Boosting uses a base learner to construct a series of weak classifiers. Each classifier is assigned a weight and when a termination criterion is reached, then the final hypothesis is constructed as a weighted combination of the classifiers from the boosting. In supervised learning, boosting uses the prediction error to determine the weights of training data and the weights for the classifiers.

Recently, boosting has been extended to semi-supervised learning. Examples are ASSEMBLE [1] and MarginBoost [3]. They use the pseudo-margin for unlabeled data in order to improve the classification performance using unlabeled examples. ASSEMBLE and MarginBoost rely only on the classifier predictions to assign the “pseudo-labels” to the unlabeled data, which is not always optimal [7]. An advantage of this approach is that it can easily be applied to multiclass classification, see [9]. A different idea is to not estimate the pseudo-margin from the predictions of the base learner. Instead, beside the margin on the labeled data, also the “consistency” is maximized. Consistency is a form of smoothness. Class labels are consistent if they are equal for data that are similar. For this the calculation of the weights for data and for classifiers in the boosting algorithm must be modified. This is done in SemiBoost [7] and RegBoost [2]. Experiments show that this idea is more effective than the approach based on pseudo-margin or confidence, see [2] and [7]. However, the solutions for the weights of data and classifiers in the boosting algorithm only exist for binary classification problems. For a multiclass semi-supervised learning problem, the only available algorithms, beside supervised learning of the labeled data only, are now those based on pseudo-margin or the use of a binary method that minimizes error on the labeled data and consistency over the unlabeled data. This must then be used in a one-versus-all or similar meta-algorithm to handle multiclass classification problems. This approach can have various problems, such as imbalanced class frequencies, increased complexity, no guarantee to obtain an optimal joint classifier, and different scales for the outputs of generated binary classifiers which complicates combining them, see [6] and [8]. Recently, in [11] a new boosting method is used for semi-supervised multiclass learning which uses similarity between predictions and data. It maps labels to n-dimensional space but this mapping may not lead to train a classifier that minimizes the margin cost properly.

In this paper we present a boosting algorithm for multiclass semi-supervised learning which minimizes both the empirical error on the labeled data and the inconsistency over labeled and unlabeled data, named Multi-SemiAdaBoost. This generalizes the SemiBoost and RegBoost algorithms from binary to multiclass classification using a coding scheme. Our proposed method uses the margin on the labeled data, the similarity among labeled and unlabeled data, and the similarity among unlabeled data in the loss function. We give a formal definition of this loss function and derive functions for the weights of classifiers and unlabeled data by minimizing an upper bound on the objective function. We then compare the performance of the algorithm to the state-of-the-art algorithms. The results show that our algorithm gives the best results.

This paper is organized as follows: section II formalizes the setting and the loss function, section III derives the weights for the boosting algorithm, sections IV and V present the experiments and the results and section VI draws the main conclusions.
II. MULTICLASS SEMI-SUPERVISED LEARNING

In this section, we define the multiclass setting and a loss function for multiclass semi-supervised learning.

A. Multiclass setting

In multiclass semi-supervised learning for the labeled points \( X_l = \{x_1, x_2, ..., x_l\} \) are provided, and for the unlabeled points \( X_u = \{x_{l+1}, x_{l+2}, ..., x_{l+u}\} \), the labels are not known. In the multiclass setting, \( Y_l \in \mathbb{R}^K \) define a K-dimensional vector with all entries equal to \(-\frac{1}{K-1}\) except a 1 in position \( i \). Each class label \( i \in \{1, ..., K\} \) is then mapped into a vector \( Y_i \). Each entry \( y_{i,j} \) is defined as follows:

\[
y_{i,j} = \begin{cases} \frac{1}{K-1} & \text{if } i = j \\ -\frac{1}{K-1} & \text{if } i \neq j \end{cases}
\]

(1)

where \( K \) is the number of classes and \( Y \) is the set of K-dimensional vectors such that for each \( Y_i \in Y \), \( \sum_{j=1}^{K} y_{i,j} = 0 \). We use the above notation to formulate the loss function for multiclass semi-supervised learning. This coding was used by Zhu et. al [12] for supervised Multiclass AdaBoost algorithm, named SAMME.

For our algorithm we also need a (symmetric) similarity matrix \( S = [S_{i,j}]_{n \times n} \), where \( S_{i,j} \) and \( S_{j,i} \) are the similarity between the points \( x_i \) and \( x_j \). \( S^l \) and \( S^u \) are the similarity matrix of labeled and unlabeled data respectively. This coding is a “meta-learner” that uses a supervised learning algorithm as base learner. We assume that the labeled and unlabeled data are drawn independently from the same data distribution. In applications of semi-supervised learning normally \( l \ll u \), where \( l \) is the number of labeled data and \( u \) is the number of unlabeled data.

B. Loss Function for Multiclass Semi-Supervised Boosting

The loss functions of the ASSEMBLE and MarginBoost algorithms for semi-supervised boosting are the sum of empirical loss on the labeled data and on the unlabeled data. Our approach is based on a different idea. Unlike ASSEMBLE and MarginBoost, instead of a term based on pseudo-margin for the unlabeled data, we add terms for consistency over the labeled and unlabeled data to the loss function. If examples that are similar, in terms of the similarity metric, are assigned different classes then this adds a penalty to the loss function. This gives three components for the loss function: (1) the empirical loss on the labeled data, (2) the consistency among labeled and unlabeled data, and (3) the consistency among unlabeled data. Combining them results in a minimization problem with the following objective function:

\[
F(Y, H) = C_1 F_l(Y, H) + C_2 F_u(Y, S^l H) + C_3 F_u(S^u H)
\]

(2)

where \( C_1, C_2, \) and \( C_3 \) are the weights for the three components. To derive the algorithm it is mathematically convenient to use an exponential loss function for margin cost on the labeled data (as in multiclass AdaBoost [12]) which gives:

\[
F_l(Y, H) = \sum_{i=1}^{n_l} \exp\left(-\frac{1}{K} Y_i H(x_i)\right)
\]

(3)

where \( H(\cdot) \) is a multiclass ensemble classifier.

In (2), \( F_u(Y, S^l H) \) measures the inconsistency between the similarity information and the classifier predictions on labeled and unlabeled data as follows:

\[
F_u(Y, S^l H) = \sum_{i=1}^{n_l} n_u \sum_{j=1}^{n_u} S^l_{i,j}(x_i, x_j) \exp\left(-\frac{1}{K} Y_i H(x_j)\right)
\]

(4)

where \( n_l \) and \( n_u \) are the number of labeled and unlabeled data respectively.

The third term of (2) measures the inconsistency between the unlabeled examples in terms of the similarity information and the classifier predictions. We use the harmonic function to define \( F_u(S^u H) \). This is a popular way to define the inconsistency in many graph-based methods, for example [13].

\[
F_u(S^u H) = \sum_{i,j \in u} S^u_{i,j}(x_i, x_j) e^{\left(\frac{1}{K} H(x_i) - H(x_j), \frac{1}{K} H(x_i)\right)}
\]

(5)

The resulting loss function is presented as the following optimization problem:

\[
\text{minimize } \quad F(Y, S, H)
\]

subject to \( H_1(x_i) + ... + H_k(x_i) = 0 \),

\[
H(x_i) = Y_i, \quad i = 1, 2, ..., n_l
\]

The above objective function is non-linear, which makes it difficult to solve. One valuable point in that function is that it uses an exponential function and since \( \exp(x) \) is convex, thus it can be shown that \( F(S, Y, H) \) maintains the convexity. Maintaining the convexity of the objective function requires that the kernel or similarity metrics be positive semi-definite. As mentioned before, \( \exp(x) \) is a convex function and the similarity function \( S(x_i, x_j) \) is non-negative for each \( i, j \). Hence, the function \( F(S, Y, H) \) is a convex function.

III. THE MULTICLASS SEMI-SUPERVISED BOOSTING ALGORITHM

In this section we derive the Multi-SemiAdaBoost algorithm. Given a set of labeled and unlabeled data, a similarity metric, and a supervised (multiclass) learning algorithm, Multi-SemiAdaBoost effectively minimizes the objective function (2). In Multi-SemiAdaBoost, a series of classifiers is constructed for the same classification task by applying a supervised base learner to varying sets of (pseudo-)labeled training data. Each classifier gets a weight.
When a termination condition is reached, the weighted combination of classifiers becomes the final hypothesis. The algorithm starts with the labeled data and in each iteration adds some unlabeled data that received a “pseudo-label” from the hypothesis constructed in the previous iteration. For this algorithm we need to find the weight for the classifiers and the confidence for the predictions. For this, we derive a criterion to assign weights to data, which are

To solve (9), we first simplify it and then with reformulation in preposition 1. We use the bound optimization approach to derive the boosting algorithm for (6), which automatically determines the step size at each iteration. We use the bound optimization technique to derive the boosting algorithm for (6), which automatically determines the step size at each iteration of boosting process.

At t-th iteration the optimization problem (2) is derived by substituting (7) in (6):

\[ F = C_1 \sum_{i=1}^{n_l} \exp \left( -\frac{1}{K} Y_i, (H_{t-1}^i(x_i) + \beta^t h^t(x_i)) \right) \]

\[ + C_2 \sum_{i=1}^{n_l} \sum_{j=1}^{n_u} S_{iu}(x_i, x_j) e^{-\frac{1}{K} Y_i, (H_{t-1}^i(x_j) + \beta^t h^t(x_j))} \]

\[ + C_3 \sum_{i,j \in n_u} S_{iu}(x_i, x_j) e^{\frac{1}{K} \sum_{k \in l} (H_{t-1}^i(x_i) + \beta^t h^t(x_j)) - (H_{t-1}^i(x_j) + \beta^t h^t(x_j))}.e_k \]

To solve (9), we first simplify it and then with reformulation we derive a criterion to assign weights to data, which are in propositions 1-2. Finally, we use proposition 3 to find optimal class labels and \( \beta \) as weight for the new classifier \( h^t(x) \). For this, we use a standard basis vector \( e_k \) to represent class membership. The resulting reformulation is in proposition 1.

**Proposition 1**: Minimizing (9) is equivalent to minimizing the following:

\[ F_1 = C_1 \sum_{i=1}^{n_l} \exp \left( -\frac{1}{K} Y_i, (H_{t-1}^i + \beta^t h^t(x_i)) \right) + C_2 \sum_{i=1}^{n_l} \sum_{j=1}^{n_u} S_{iu}(x_i, x_j) \exp \left( -\frac{1}{K} Y_i, e_k \right) \]

\[ + C_3 \sum_{i,j \in n_u} S_{iu}(x_i, x_j) \exp \left( \frac{1}{K} (H_{t-1}^i - H_{t-1}^j).e_k \right) \]

\[ \exp \left( \frac{\beta^t}{K} (h^t_i - h^t_j).e_k \right) \]

Subject to:

\[ H_{t}^i(x_i) + ... + H_{t}^i(x_i) = 0 \]

\[ H_{t}^i(x_i) = Y_i \text{ for } i = 1, ..., n_l \]

where \( \delta(Y_i, e_k) \) is:

\[ \delta(Y_i, e_k) = \begin{cases} 1 & \text{if } i = k \\ 0 & \text{if } i \neq k \end{cases} \]

**Proof**: To simplify the notation we write \( H_{t-1}^i \) for \( H_{t-1}^i(x) \) and \( h^t \) for \( h^t(x) \), see appendix.

For now assume that \( C_1 = C_2 = C_3 = 1 \). As mentioned before, (10) is a nonlinear optimization problem. To solve this problem and also to find a criterion for assigning “pseudo-labels” to the unlabeled examples we use Proposition 2.

**Proposition 2**: Minimizing \( F_1 \) is equivalent to minimizing the following objective function \( F_2 \):

\[ F_2 = \sum_{i=1}^{n_l} W_i \exp \left( -\frac{1}{K} Y_i, h^t_i \right) + \sum_{i=1}^{n_u} \sum_{k} P_{i,k} \exp \left( -\frac{1}{K} h^t_i.e_k \right) \]

where \( W_i = \exp \left( -\frac{1}{K} Y_i, H_{t-1}^i \right) \)

and

\[ P_{i,k} = \sum_{j=1}^{n_l} S_{iu}(x_i, x_j) e^{\frac{1}{K} (H_{t-1}^i - H_{t-1}^j).e_k} \delta(Y_i, e_k) \]

\[ + \sum_{j=1}^{n_u} S_{iu}(x_i, x_j) e^{\frac{1}{K} (H_{t-1}^i - H_{t-1}^j).e_k} e^{\frac{1}{K}} \]

**Proof**: See appendix.

In the optimization problem (12), \( P_{i,k} \) can be used for weighting the unlabeled examples and \( W_i \) is used for weighting labeled data. According to these criteria the algorithm decides whether to select a sample for the new training set or not. The expression in (12) is in terms of \( \beta^t \) and \( h^t(x) \) and hence make it difficult to solve. The following proposition simplifies it by using a bound.
Proposition 3: The class label for unlabeled example \( x_i \) that minimizes \( F_2 \) is:

\[
\hat{Y}_i = \arg \max_k (P_{t,k})
\]  

and the weight for that will be \( w'_i = \max |P_{t,k}| \). Meanwhile the optimal value for \( \beta \) that minimizes \( F_2 \) is:

\[
\beta = \frac{(K - 1)^2}{K} \left( \log(K - 1) + \log\left(1 - \frac{\epsilon^t}{\epsilon^t - \epsilon^t}ight) \right)
\]

where \( P_{t} \equiv P(x_i) \) was defined in (7).

**Proof:** See appendix.

Now, let \( \epsilon^t \) denote the weighted error made by the classifier \( h^t \). Then the \( \epsilon^t \) will be as follows:

\[
\epsilon^t = \frac{\sum_{i \in n_u} \sum_{k \in \ell} P_{t,k} \delta(h^t_i, e_k, P_t = k) + \sum_{i \in n_l} W_i}{\sum_{i \in n_u} \sum_{k \in \ell} P_{t,k} \delta(h^t_i, e_k, P_t \neq k) + \sum_{i \in n_l} W_i}
\]

Replacing (17) in (16) leads to the following \( \beta^t \):

\[
\beta^t = \frac{(K - 1)^2}{K} \left( \log(K - 1) + \log\left(\frac{1 - \epsilon^t}{\epsilon^t - \epsilon^t}\right) \right)
\]

which is a generalization of the weighting factor of Multiclass AdaBoost [12]. Our formulation can thus be seen as a generalization of AdaBoost to multiclass semi-supervised data.

A. Multi-SemiAdaBoost Algorithm

Based on the previous discussion, we can provide the details of the algorithm. In Algorithm 1, first the values of (SAMME). Then (15) is used to assign a “pseudo-label” to the unlabeled example and the value of (15) becomes the weight. Next, a set of high-confidence newly-labeled data besides the labeled data are used based on the weights for training a new classifier, called \( h^t(x) \). The algorithm then uses the result of Proposition (2) to sample data which will lead to a decrease of the value of the objective function. This new set is consistent with two criteria, classifier prediction and similarity between examples. The boosting process is repeated until it reaches a stopping condition. After finishing, the final hypothesis is the weighted combination of the generated classifiers. We use \( \beta^t \leq 0 \) and a fix number of iterations as stopping conditions.

**Proposition 3:** The class label for unlabeled example \( x_i \) that minimizes \( F_2 \) is:

\[
\hat{Y}_i = \arg \max_k (P_{t,k})
\]

where \( P_{t} \equiv P(x_i) \) was defined in (7).

**Proof:** See appendix.

Now, let \( \epsilon^t \) denote the weighted error made by the classifier \( h^t \). Then the \( \epsilon^t \) will be as follows:

\[
\epsilon^t = \frac{\sum_{i \in n_u} \sum_{k \in \ell} P_{t,k} \delta(h^t_i, e_k, P_t = k) + \sum_{i \in n_l} W_i}{\sum_{i \in n_u} \sum_{k \in \ell} P_{t,k} \delta(h^t_i, e_k, P_t \neq k) + \sum_{i \in n_l} W_i}
\]

Replacing (17) in (16) leads to the following \( \beta^t \):

\[
\beta^t = \frac{(K - 1)^2}{K} \left( \log(K - 1) + \log\left(\frac{1 - \epsilon^t}{\epsilon^t - \epsilon^t}\right) \right)
\]

which is a generalization of the weighting factor of Multiclass AdaBoost [12]. Our formulation can thus be seen as a generalization of AdaBoost to multiclass semi-supervised data.

**A. Multi-SemiAdaBoost Algorithm**

Based on the previous discussion, we can provide the details of the algorithm. In Algorithm 1, first the values of (SAMME). Then (15) is used to assign a “pseudo-label” to the unlabeled example and the value of (15) becomes the weight. Next, a set of high-confidence newly-labeled data besides the labeled data are used based on the weights for training a new classifier, called \( h^t(x) \). The algorithm then uses the result of Proposition (2) to sample data which will lead to a decrease of the value of the objective function. This new set is consistent with two criteria, classifier prediction and similarity between examples. The boosting process is repeated until it reaches a stopping condition. After finishing, the final hypothesis is the weighted combination of the generated classifiers. We use \( \beta^t \leq 0 \) and a fix number of iterations as stopping conditions.

**IV. EXPERIMENTS**

In the experiments we compare Multi-SemiAdaBoost (MSAB) with several other algorithms. One comparison is with using the base learner on the labeled data only and a second is with (multiclass) AdaBoost (SAMME [12]) using the same base learner. The purpose is to evaluate if MSAB exploits the information in the unlabeled data. We also include comparisons with the state-of-the-art algorithms for semi-supervised boosting, in particular ASSEMBLE, SemiMultiAdaBoost [9], RegBoost [2], and SemiBoost [7]. Like MSAB, SemiBoost and RegBoost use smoothness regularization but they are limited to binary classification.

For comparison, we use the one-vs-all method to handle the multiclass classification problem with RegBoost and SemiBoost. In our experiments, we use WEKA implementation of the classifiers with default parameter settings [5].

Two main steps in Multi-SemiAdaBoost are: (i) the similarity between examples and (ii) sample from the unlabeled examples. There are different distance-based approaches to compute the similarity between the data. In this paper we employ the Radial Basis Function (RBF) which is used effectively as similarity metric in many domains. For two examples \( x_i \) and \( x_j \) the RBF similarity is computed as:

\[
S(x_i, x_j) = \exp\left(-\frac{||x_i - x_j||^2}{\sigma^2}\right)
\]

where \( \sigma \) is the scale parameter which plays a major role in the performance of the kernel, and should be carefully tuned to the problem [13].

In the sampling process only the high-confidence data points must be selected for training. Finding the best selection is a difficult problem [10]. On one hand, selecting a small set of newly-labeled examples might lead to slow convergence, and on the other hand, selecting a large set of newly-labeled examples may include some poor examples. One possible solution for that is to use a threshold or even a fixed number which is optimized through the training process. Sampling data for labeling is based on the following
We use amounts to a measure of confidence. For the labeled data probability distribution of classes of the example \( x_i \) reported refer to the test set. With different subsets of training and testing data, the results to the original data set. We run each experiment 10 times labeled data, keeping the class proportions in all sets similar partitioned into 90 percent unlabeled data and 10 percent as training data. Training data in each experiment are first repository are used in our experiments. For each dataset, 30 \([7], \) and \([9]\). Sixteen benchmark datasets from the UCI data used for evaluating semi-supervised learning methods \([10]\), algorithms. Recently several UCI datasets have been extensively
tree classifier) and Naive Bayes as base classifiers. The resulting algorithm is shown to perform better than existing algorithms for multiclass problems. Multi-SemiAdaBoost significantly improves the performance of supervised learning with different base classifiers for nearly all the datasets. Using statistical t-test, we observed that Multi-SemiAdaBoost improves the performance of J48 and Naive Bayes base classifiers on 16 out of 16 datasets. The results show that Multi-SemiAdaBoost is also better than multiclass AdaBoost meta classifier trained using only labeled data. We also observe that the classification models generated by using Multi-SemiAdaBoost are relatively more stable than J48 base classifier because of lower standard deviation in classification accuracy.

**MSAB and Supervised Learning**

Multi-SemiAdaBoost significantly improves the performance of supervised learning with different base classifiers for nearly all the datasets. Using statistical t-test, we observed that Multi-SemiAdaBoost improves the performance of J48 and Naive Bayes base classifiers on 16 out of 16 datasets. The results show that Multi-SemiAdaBoost is also better than multiclass AdaBoost meta classifier trained using only labeled data. We also observe that the classification models generated by using Multi-SemiAdaBoost are relatively more stable than J48 base classifier because of lower standard deviation in classification accuracy.

**MSAB and Semi-Supervised Boosting Algorithms**

Table I and II show that Multi-SemiAdaBoost gives better results than all four semi-supervised algorithms ASSEMBLE, SemiMultiAdaBoost, RegBoost, and SemiBoost on 14 (for J48) or 13 (for Naive Bayes) out of 16 datasets. The improvement of MASB in most of the used datasets are significant and it outperforms the ASSEMBLE and SBoost on 15 out of 16 datasets, when the base classifier is J48. The same results can be seen with Naive Bayes classifier.

There are datasets where the proposed algorithms may not significantly improve the performance of the base classifiers. In these cases the supervised algorithm outperforms all the semi-supervised algorithms, for example in Tables II the AdaBoost meta classifier performs the same as the proposed methods on Car, Cmc, and Diabetes datasets and outperforms the other methods. These kinds of results emphasize that the unlabeled examples do not guarantee that they always are useful and improve the performance.

**VI. Conclusion and Discussion**

We presented a boosting algorithm for multiclass semi-supervised learning with a novel loss function involving the empirical error on labeled data and the consistency between classifier predictions and similarity information. Specifically we derived the weights for the data and for the classifiers. The resulting algorithm is shown to perform better than existing algorithms for multiclass problems. Multi-SemiAdaBoost can use any multiclass supervised learning algorithm as base classifier.

An open problem for future work is how to efficiently find and tune a good similarity function. Another issue is the encoding for the multiclass setting. The encoding that is used here is not the only possibility and it seems interesting to exploit other encoding schemes.

**References**

Table I: The classification accuracy (± s.d) with 10% labeled examples and J48 as base learner

<table>
<thead>
<tr>
<th>DataSets</th>
<th>Supervised Learning</th>
<th>Semi-Supervised Learning</th>
<th>ASSEMBLE</th>
<th>SMOBoost</th>
<th>RegBoost</th>
<th>SemiBoost</th>
<th>MSAB</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DT SAMME</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Balance(625,3)</td>
<td>70.72±5.5</td>
<td>74.75±3.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Car(1728,4)</td>
<td>76.93±1.6</td>
<td>78.34±2.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cmc(1473,3)</td>
<td>42.03±3.8</td>
<td>44.89±4.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dermatology(366,6)</td>
<td>73.00±4.1</td>
<td>76.99±5.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diabetes(68,2)</td>
<td>68.81±5.0</td>
<td>70.07±5.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ecoli(136,8)</td>
<td>65.88±9.3</td>
<td>66.82±8.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Glass(192,6)</td>
<td>49.79±9.6</td>
<td>50.94±6.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Iris(50,3)</td>
<td>71.13±8.9</td>
<td>71.13±8.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Liver(345,2)</td>
<td>54.86±7.5</td>
<td>56.93±7.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Optdigits(1409,10)</td>
<td>63.26±2.4</td>
<td>67.52±3.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sonar(208,2)</td>
<td>60.08±9.2</td>
<td>59.87±6.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Soybean(880,19)</td>
<td>51.00±4.4</td>
<td>47.85±4.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vowel(990,11)</td>
<td>55.92±4.7</td>
<td>52.57±3.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wave(500,3)</td>
<td>70.10±1.7</td>
<td>78.04±8.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wine(178,3)</td>
<td>67.36±9.8</td>
<td>67.36±9.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zoo(101,7)</td>
<td>66.19±5.2</td>
<td>56.66±8.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table II: The classification accuracy (± s.d) with 10% labeled examples and Naive Bayes as base learner

<table>
<thead>
<tr>
<th>DataSets</th>
<th>Supervised Learning</th>
<th>Semi-Supervised Learning</th>
<th>ASSEMBLE</th>
<th>SMOBoost</th>
<th>RegBoost</th>
<th>SemiBoost</th>
<th>MSAB</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Naive Bayes SAMME</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Balance</td>
<td>75.88±3.5</td>
<td>76.71±2.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Car</td>
<td>77.5±1.9</td>
<td>78.51±0.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cmc</td>
<td>46.98±1.4</td>
<td>48.05±1.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dermatology</td>
<td>81.52±2.1</td>
<td>81.52±2.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diabetes</td>
<td>74.35±3.3</td>
<td>76.99±2.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ecoli</td>
<td>81.68±3.3</td>
<td>78.69±6.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Glass</td>
<td>50.7±6.7</td>
<td>50.21±6.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Iris</td>
<td>81.25±5.5</td>
<td>80.95±6.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Liver</td>
<td>55.39±5.5</td>
<td>55.04±7.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Optdigits</td>
<td>72.04±1.9</td>
<td>70.57±1.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sonar</td>
<td>64.41±7.1</td>
<td>63.33±5.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Soybean</td>
<td>70.00±1.4</td>
<td>68.44±7.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vowel</td>
<td>47.52±3.4</td>
<td>48.02±3.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wave</td>
<td>80.44±1.4</td>
<td>80.91±1.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wine</td>
<td>84.56±5.8</td>
<td>84.56±5.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zoo</td>
<td>87.77±4.0</td>
<td>87.77±4.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>


APPENDIX

Proposition 1:
Proof: Let the function $Y_i, e_k$ be equal to 1, if a data point with $Y_i$ belongs to class $k$. Then (9) can be rewritten as follows:

$$\min F = C_1 \sum_{i=1}^{n_i} \exp \left( - \frac{1}{K} Y_i (H_{t-1}^i(x_i) + \beta^t h^i_t(x_i)) \right)$$
$$+ C_2 \sum_{i=1}^{n_i} \sum_{j=1}^{n} S^{lu}(x_i, x_j) e^{- \frac{1}{K} \sum_{k \in \{i, e_k\}} (H_{j-1}^i + h^i_t) e_k}$$
$$+ C_3 \sum_{i,j \in n} S^{nu}(x_i, x_j)$$
$$e^{\frac{1}{K} \sum_{k \in \{i, e_k\}} \left( (H_{j-1}^i + h^i_t) - (H_{j-1}^i + h^i_t) \right) e_k}$$

Using the following inequality in (9):

$$(c_1 c_2 \ldots c_n)^\frac{1}{n} \leq \frac{c_1 + c_2 + \ldots + c_n}{n}$$

and then replacing the value of $Y_i$ results in the (10), where $\delta$ is defined in (11).

Proposition 2:
Proof: According to the formulation that we mentioned earlier, $\frac{1}{\beta^t} \leq \beta^t h(x_i).e_k \leq 1$, then multiplying it with $\frac{1}{K-1}$ and using the exponential function gives the following inequality:

$$e^{\frac{1}{K-1}} \leq e^{\frac{\beta^t}{K-1} h(x_i)} \leq e^{\frac{1}{K-1}}$$

Using (24) for decomposing the third term of (10) leads to the following expression:

$$e^{\frac{\beta^t}{K-1} (h(x_i) - h(x_j)) e_k} \leq e^{\frac{1}{K-1}} e^{-\frac{\beta^t}{K-1} h(x_j) e_k}$$

Replacing the inequality (25) in the last term of (10) gives:

$$F_1 = \sum_{i=1}^{n_i} \exp \left( - \frac{1}{K} Y_i H_{t-1}^i \right) \exp \left( - \frac{\beta^t}{K} Y_i h^i_t \right)$$
$$+ \sum_{i=1}^{n_i} \sum_{j=1}^{n} \sum_{k \in \{i, e_k\}} S^{lu}(x_i, x_j) \exp \left( \frac{1}{K-1} H_{j-1}^i e_k \right)$$
$$\exp \left( - \frac{\beta^t}{K-1} h^i_t e_k \right) \delta \left( Y_i, e_k \right)$$
$$+ \sum_{i,j \in n} \sum_{k \in \{i, e_k\}} S^{nu}(x_i, x_j) \exp \left( \frac{1}{K-1} (H_{j-1}^i - H_{t-1}^i) e_k \right)$$
$$\left( e^{\frac{1}{K-1}} e^{-\frac{\beta^t}{K-1} h^i_t e_k} \right)$$

Factoring out the common term $e^{\frac{\beta^t}{K-1} h(x_j) e_k}$ and re-arranging terms then results in:

$$F_1 \leq \sum_{i=1}^{n_i} \exp \left( \frac{1}{K} Y_i H_{t-1}^i \right) \exp \left( - \frac{\beta^t}{K} Y_i h^i_t \right)$$
$$+ \sum_{j=1}^{n} \sum_{k \in \{i, e_k\}} S^{lu}(x_i, x_j) \exp \left( \frac{1}{K-1} H_{j-1}^i e_k \right)$$
$$\exp \left( - \frac{\beta^t}{K-1} h^i_t e_k \right) \delta \left( Y_i, e_k \right)$$
$$+ \sum_{i,j \in n} \sum_{k \in \{i, e_k\}} S^{nu}(x_i, x_j) \exp \left( \frac{1}{K-1} (H_{j-1}^i - H_{t-1}^i) e_k \right)$$
$$\left( e^{\frac{1}{K-1}} e^{-\frac{\beta^t}{K-1} h^i_t e_k} \right)$$

As a result, it gives:

$$F_1 \leq \sum_{i=1}^{n_i} W_i \exp \left( \frac{-\beta^t}{K} Y_i h^i_t \right) + \sum_{i=1}^{n_i} \sum_{k \in \{i, e_k\}} P_{i,k} \exp \left( - \frac{\beta^t}{K-1} h^i_t e_k \right)$$

where

$$W_i = \exp \left( \frac{-1}{K} Y_i H_{t-1}^i \right)$$

and

$$P_{i,k} = \sum_{j=1}^{n} S^{lu}(x_i, x_j) \exp \left( \frac{1}{K-1} (H_{j-1}^i - H_{t-1}^i) e_k \right)$$

Proposition 3:
Proof: We use the following inequality to decompose the elements of $F_2$:

$$\forall x \in [-1, 1] \quad \exp(\beta x) \leq \exp(\beta) + \exp(-\beta) + \beta x - 1,$$

Replacing the inequality (31) in (12) gives:

$$F_2 \leq \sum_{i=1}^{n_i} W_i \exp \left( \frac{\beta^t}{K} + \frac{\beta^t}{K} - 1 \right) - \sum_{i=1}^{n_i} \sum_{k \in \{i, e_k\}} P_{i,k} \exp \left( \frac{-\beta^t}{K-1} h^i_t e_k \right)$$

Then, re-arranging the above inequality gives:

$$F_2 \leq \left( \sum_{i=1}^{n_i} W_i \exp \left( \frac{\beta^t}{K} + \frac{\beta^t}{K} - 1 \right) - \sum_{i=1}^{n_i} \sum_{k \in \{i, e_k\}} P_{i,k} \exp \left( \frac{-\beta^t}{K-1} h^i_t e_k \right) \right)$$

The first term in inequality (33) is independent of $h_i$. Hence, to minimize $F_2$, finding the examples with the largest $P_{i,k}$ and $W_i$ is sufficient at each iteration of the boosting process.
We assign the value $P_{i,k}$ of selected examples as weight for the newly-labeled examples, hence $w'_{i} = \max |P_{i,k}|$. To prove the second part of the proposition, we expand $F_2$ as follows:

\[
F_2 = \sum_{i=1}^{n_l} W_i \exp\left(\frac{-\beta^t}{K} Y_i h^t_i\right) + \sum_{i=1}^{n_u} \sum_{k \in L} P_{i,k} \exp\left(\frac{-\beta^t}{K-1} h^t_i e_k\right)
\]

\[
= \sum_{i \in n_l, h^t_i = Y_i} W_i \exp\left(\frac{-\beta^t}{K-1}\right) + \sum_{i \in n_l, h^t_i \neq Y_i} \sum_{k \in L} P_{i,k} \exp\left(\frac{-\beta^t}{(K-1)^2}\right) + \sum_{i=1}^{n_u} \sum_{k \in L} P_{i,k} \exp\left(\frac{\beta^t}{(K-1)^2}\right) \delta'(h^t_i e_k, P_i = k) + \sum_{i=1}^{n_u} \sum_{k \in L} P_{i,k} \exp\left(\frac{-\beta^t}{(K-1)^2}\right) \delta'(h^t_i e_k, P_i \neq k)
\]

(34)

where $\delta'$ is defined as:

\[
\delta'(h^t_i e_k, T) = \begin{cases} 1 & \text{if } T \text{ is true} \\ 0 & \text{otherwise} \end{cases}
\]

(35)

Differentiating the above equation w.r.t $\beta$ and equating it to zero, gives:

\[
\frac{\partial F_2}{\partial \beta} = -\frac{1}{K-1} \exp\left(\frac{-\beta}{K-1}\right) \sum_{i \in n_l, h^t_i = Y_i} W_i + \frac{1}{(K-1)^2} \exp\left(\frac{\beta}{(K-1)^2}\right) \sum_{i \in n_l, h^t_i \neq Y_i} W_i
\]

\[
+ \frac{1}{K-1} \exp\left(\frac{-\beta}{K-1}\right) \sum_{i=1}^{n_u} \sum_{k \in L} P_{i,k} \delta'(h^t_i e_k, P_i = k) + \frac{1}{(K-1)^2} \exp\left(\frac{\beta}{(K-1)^2}\right) \sum_{i=1}^{n_u} \sum_{k \in L} P_{i,k} \delta'(h^t_i e_k, P_i \neq k) = 0
\]

(36)

Simplifying the above equation results in (16).