ON IMPROVING SEMI-SUPERVISED MARGINBOOST INCREMENTALLY USING STRONG UNLABELED DATA*

Thanh-Binh Le and Sang-Woon Kim
Department of Computer Engineering, Myongji University, 449-728, Yongin, South Korea

Keywords: Semi-supervised MarginBoost, Incremental learning strategy, Dissimilarity-based classifications.

Abstract: The aim of this paper is to present an incremental learning strategy by which the classification accuracy of the semi-supervised MarginBoost (SSMB) algorithm (d'Alché Buc, 2002) can be improved. In SSMB, both a limited number of labeled and a multitude of unlabeled data are utilized to learn a classification model. However, it is also well known that the utilization of the unlabeled data is not always helpful for semi-supervised learning algorithms. To address this concern when dealing with SSMB, in this paper we study a means of selecting only "small" helpful portion of samples from the additional available data. More specifically, this is done by performing SSMB after incrementally reinforcing the given labeled training data with a part of strong unlabeled data; we train the classification model in an incremental fashion by employing a small amount of "strong" samples selected from the unlabeled data per iteration. The proposed scheme is evaluated with well-known benchmark databases, including some UCI data sets, in two approaches: dissimilarity-based classification (DBC) (Pekalska and Duin, 2005) as well as conventional feature-based classification. Our experimental results demonstrate that, compared to previous approaches, it achieves better classification accuracy results.

1 INTRODUCTION

MarginBoost (Mason, 2000) aims at improving the classification performance of an ensemble classifier designed with weak classifiers by means of linear combination. By introducing a means of generating the MarginBoost in a semi-supervised approach, semi-supervised MarginBoost (SSMB) was proposed (d'Alché Buc, 2002). In SSMB, a large amount of unlabeled data, U, together with labeled data, L, are used to build better classifiers. That is, the algorithm exploits the samples of U in addition to the labeled counterparts to improve the performance on a classification task, leading to a performance improvement of the supervised learning algorithm with a multitude of unlabeled data.

However, it is also well known that U does not always help during SSMB learning. Specifically, it is not guaranteed that adding U to the training data, T, i.e., $T = L \cup U$, leads to a situation in which we can improve the classification performance. Therefore, if we can know more about confidence levels involved in classifying U, we could choose some of the informative data and include it when training weak classifiers. This idea has been used in SemiBoost (Mallapragada, 2009), where the authors measured the pairwise similarity to guide the selection of a subset of U at each iteration and to assign labels to them.

To improve the performance of SSMB further, in this paper we propose a modified SSMB algorithm in which we use the discriminating unlabeled data in an *incremental* fashion rather than in batch mode (Cesa-Bianchi, 2006). In both SemiBoost and the modified SSMB, some instances of the strong unlabeled data are selected from the given U and are then used to train the classification model in addition to L. However, the two algorithms differ in how they construct T. In the present SSMB, the cardinality of T is increased incrementally as the iterations are repeated, while, in SemiBoost, the cardinality of T is always the same when executing the learning iterations.

The main contribution of this paper is that it demonstrates that the classification accuracy of SSMB can be improved by incrementally utilizing a portion of the unlabeled data as well as the labeled training data. Also, an evaluation of the proposed scheme has been performed in two fashions: traditional feature-based classification (Fukunaga, 1990) and recently developed dissimilarity-based classification (Pekalska and Duin, 2005).

*This work was supported by the National Research Foundation of Korea funded by the Korean Government (NRF-2011-0002517).

Le T. and Kim S. (2012).
ON IMPROVING SEMI-SUPERVISED MARGINBOOST INCREMENTALLY USING STRONG UNLABELED DATA.
In *Proceedings of the 1st International Conference on Pattern Recognition Applications and Methods*, pages 265-268
DOI: 10.5220/0003721202650268

2 SSMB IMPROVED

In SSMB, an ensemble classifier, g_t, is designed with weak classifiers, $h_\tau \in \mathcal{H}$, by means of a linear combination, as follows: $g_t(x) = \sum_{\tau=1}^{t} \alpha_\tau h_\tau(x)$, where α_τ is a normalized step-length. For the training data, $T = L \cup U$, where $L = \{(x_i, y_i)\}_{i=1}^{n_l}$ and $U = \{(x_j)\}_{j=1}^{n_u}$, the algorithm minimizes the cost function C defined with any scalar decreasing function c of the margin ρ: $C(g_t) = \sum_{i=1}^{n_l} c(\rho_L(g_t(x_i), y_i)) + \sum_{i=1}^{n_u} c(\rho_U(g_t(x_i)))$, where $\rho_L(g_t(x_i), y_i) = y_i g_t(x_i)$ and $\rho_U(g_t(x_i)) = g_t(x_i)^2$. Here, the criterion quantities for L and U, J_t^L and J_t^U, are expressed as:

$$J_t^L = \sum_{x_i \in L} w_t(i) y_i h_{t+1}(x_i), \quad (1)$$

$$J_t^U = \sum_{x_i \in U} w_t(i) \frac{\partial \rho_U(g_t(x_i))}{\partial g_t(x_i)} h_{t+1}(x_i), \quad (2)$$

where $w_t(i)$ is computed as follows:

$$w_t(i) = \begin{cases} \frac{c'(\rho_L(g_t(x_i), y_i))}{\sum_{x_j \in T} w_{t-1}(j)}, & \text{if } x_i \in L, \\ \frac{c'(\rho_U(g_t(x_i)))}{\sum_{x_j \in T} w_{t-1}(j)}, & \text{if } x_i \in U. \end{cases} \quad (3)$$

An algorithm for SSMB is formalized as follows:

1. Initialization: $g_0(x) = 0$; $w_0(i) = \frac{1}{n_l + n_u}, i = 1, \cdots, n_l + n_u$.

2. Compute predicted labels of U using the nearest neighbor (NN) rule.

3. Do the following steps while increasing t by unity from 1 to t_1 per epoch:

(a) Learn the gradient direction h_t for T while maximizing J_t^T $(= J_t^L + J_t^U)$ computed with (1, 2).

(b) If $J_t^T \leq 0$, then exit and return $g_t(x)$; otherwise, go to the next sub-step.

(c) After computing $g_{t+1}(x) = g_t(x) + \alpha_t h_t(x)$, update the weights $w_{t+1}(i)$ with (3) for the next iteration.

As mentioned previously, the unlabeled data do not always help in SSMB learning processes. In particular, when the cardinality of the unlabeled data is much larger than that of labeled data, the situation is much worse. To overcome the limitation based on this, we expand SSMB using classification confidence of the unlabeled data as SemiBoost does. The present SSMB and SemiBoost select the strong examples from unlabeled data based on the confidence level. However, two algorithms differ in terms of the following points: how they select the strong samples from the unlabeled data and how they determine pseudo-labels of the selected unlabeled data. In SemiBoost, 10% of the entire unlabeled data set is repeatedly selected based on the confidence levels, while in the present SSMB, 10% of the currently available unlabeled data is selected incrementally. Also, the two

Figure 1: A comparison of the training data sets of SSMB, improved SSMB, and SemiBoost learning algorithms.

algorithms are different in how they determine the class labels of the selected unlabeled data. The former determines the pseudo-labels based on the similarity matrix, but the latter determines them based on the NN rule. On the basis of what we have briefly discussed, an algorithm for the present SSMB is formalized as follows:

1. This step is the same as Step 1 in SSMB.

2. For all $x_i, x_j \in T$, compute a similarity matrix, $s(i, j) = exp(-\|x_i - x_j\|_2^2 / \sigma^2)$, where σ is a scale parameter, and predicted labels of U using NN rule.

3. Repeat the following steps while increasing t by unity from 1 to t_1 per epoch: First, using a "sampling", obtain new training data, T_t, from available unlabeled data, T_u, in addition to L. Next, for T_t, repeat the three sub-steps of Step 3 of SSMB ten times.

In the above sampling process, we first choose a portion of the strong data from T_u (i.e., 10%; T_u^{10}) according to the confidence levels based on s. Then, we update $T_u \leftarrow T_u - T_u^{10}$ and $n_u \leftarrow |T_u|$. Fig.1 shows a comparison of the training data set T_t of SSMB, the modified (and improved) SSMB, and SemiBoost.

3 EXPERIMENTAL RESULTS

The proposed scheme was tested and compared with conventional methods. This was done by performing experiments on the well-known benchmark databases of Nist389, mfeat-fac, and mfeat-kar, as well as other multivariate data sets cited from the UCI Machine Learning Repository [2].

In this experiment, the data sets are initially split into three parts: labeled training sets, labeled test sets, and unlabeled data at a ratio of 20 : 10 : 70. The training and test procedures are then repeated ten times and the results obtained are averaged. Specifically, the classifications are performed in two fashions: feature-based classification (FBC) and dissimilarity-based classification (DBC). In DBC, the classifica-

[2] http://www.ics.uci.edu/~mlearn/MLRepository.html.

tion process is not based on the feature measurements of individual object samples but rather on a suitable dissimilarity measure among the individual samples. Therefore, in this experiment, after measuring the dissimilarity among paired samples with the Euclidean distance, the classifications were performed on the constructed dissimilarity matrix. In the interest of compactness, the details of DBC are omitted here, but can be found in (Pekalska and Duin, 2005).

Conventional SSMB and the newly proposed SSMB (which are referred to as SSMB-original and SSMB-improved, respectively) were performed with numbers of weak learners ranging from 10 to 50 in increments of 5 at a time. This was repeated ten times. The scalar decreasing function employed for the margin ρ was $c(x) = e^{-x}$. In particular, the step-length $\alpha_t = \frac{1}{4}\ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)$, where $\varepsilon_t = \sum_i w_t(i)\delta(y_i g_t(x_i), -1)$ was commonly used for both SSMBs. For all of the boosting algorithms, a decision-tree classifier was used as the weak learner and implemented with Prtools (Duin and Tax, 2004).

First, the experimental results obtained with the two classifying approaches, FBC and DBC, for the Nist389, mfeat-fac, and mfeat-kar databases were investigated. Fig. 2 shows a comparison of the error rates (and standard deviations) of SemiBoost, SSMB-original, and SSMB-improved, obtained with the two classifying approaches for Nist389. Here, in the interest of brevity, the other results are omitted. Also, to reduce the computational complexity, the dimensionality of all of the data sets was reduced to 10 values using a principal component analysis (PCA).

From the figures shown in Fig. 2, it can be observed that, in general, the classification accuracies of SSMB, estimated with FBC and DBC, can be improved. This is clearly shown in the error rates of the ensemble classifiers, as represented by the red lines (dashed and solid lines with the $\diamondsuit$ marker) and the black and blue lines (dashed and solid lines with the $\square$ and $\ominus$ markers, respectively). For all three data sets and for each repetition, the rank of achieving the lowest error rate is always identical in the order of SSMB-improved, SSMB-original, and SemiBoost. That is, the winner is always the SSMB-improved. In addition, it should be pointed out that the improvements of the two methods of DBC and FBC were similar. According to the different number of repetitions, the increase and/or decrease in the error rates of the two approaches appeared to be consistent.

Furthermore, the following is an interesting issue to investigate: *Is the classification accuracy of the improved SSMB algorithm better (or more robust) than those of conventional schemes when changing the amount of the selected strong data?* To answer this question, for the data sets, we repeated the experiment with *four* different strong data sizes (i.e., $T_u^5, T_u^{10}, T_u^{15}, T_u^{20}$) and ten repetitions, as was done previously under the same experimental conditions. The experimental results in this case showed that the error rates obtained with the four differently sized instances of strong unlabeled data are similar.

Table 1 shows a numerical comparison of the error rates obtained with AdaBoost, MarginBoost, SemiBoost, and the original and improved SSMB algorithms for the three data sets. Here, two supervised boosting algorithms, AdaBoost and MarginBoost, were employed as a reference for comparison. These supervised algorithms were trained with only 20% of the labeled training data and were evaluated with 10% of the labeled test data, while the three semi-supervised algorithms, SemiBoost, SSMB-original, and SSMB-improved, were trained with 70% of the unlabeled training data as well as 20% of the labeled data. They were also evaluated also with 10% of the labeled test data. For all of the boosting algorithms, the number of weak classifiers was identical, at $t_1 = 50$. In the table, the estimated error rates that increase and/or decrease more than the sum of the standard deviations are underlined.

To investigate the advantage of incrementally using strong unlabeled data further and especially to determine which types of significant data sets are more suitable for the scheme, we repeated the experiment with a few UCI data sets. From the results obtained, as in Table 1, it should be noted again that the classification accuracy of the SSMB algorithm can be generally improved when utilizing the unlabeled data in an incremental learning fashion. However, the proposed scheme does not work satisfactorily with *low*-dimensional data sets. That is, for *high*-dimensional data sets, the difference in the error rates of SSMB-original and SSMB-improved schemes is relatively large, whereas the difference in the error rates for *low*-dimensional data sets is marginal.

4 CONCLUSIONS

In an effort to improve the classification performance of SSMB, in this paper we used an incremental learning strategy with which the SSMB can be implemented efficiently. We first computed the similarity matrix of labeled and unlabeled data and, in turn, selected a small amount of strong unlabeled samples based on their confidence levels. We then trained a classification model using the selected unlabeled samples as well as labeled samples in an incremental fashion. The proposed strategy was evaluated with well-

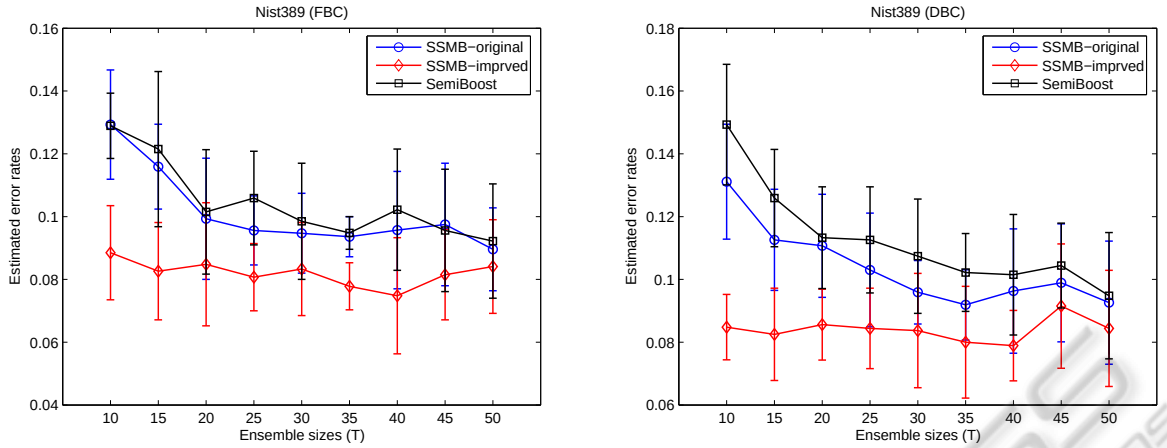

Figure 2: A comparison of the estimated error rates (standard deviations) obtained with the FBC and DBC approaches for Nist389: (a) left and (b) right; (a) and (b) are of FBC and DBC, obtained with SemiBoost and the two SSMB algorithms.

Table 1: A numerical comparison of the error rates (standard deviations) obtained with two supervised schemes (i.e., *AdaBoost* and *MarginBoost*) and three semi-supervised schemes (i.e., *SemiBoost*, *SSMB-original*, and *SSMB-improved*) for the three data. Here, three numbers in brackets of the first column represent the dimensions d, samples n, and classes c, respectively.

data sets (d/n/c)	classifier types	supervised learning		semi-supervised learning		
		AdaBoost	MarginBoost	SemiBoost	SSMB-original	SSMB-imprved
Nist389	FBC	0.0648(0.0130)	0.0648(0.0130)	0.0730(0.0155)	0.0696(0.0142)	0.0496(0.0122)
(1024/1500/3)	DBC	0.0563(0.0134)	0.0537(0.0139)	0.0574(0.0129)	0.0652(0.0165)	0.0400(0.0110)
mfeat-fac	FBC	0.0131(0.0033)	0.0131(0.0036)	0.0136(0.0038)	0.0156(0.0033)	0.0099(0.0022)
(216/2000/10)	DBC	0.0258(0.0035)	0.0258(0.0034)	0.0270(0.0054)	0.0280(0.0044)	0.0227(0.0046)
mfeat-kar	FBC	0.0236(0.0025)	0.0236(0.0025)	0.0234(0.0033)	0.0224(0.0030)	0.0166(0.0027)
(64/2000/10)	DBC	0.0167(0.0029)	0.0166(0.0029)	0.0175(0.0025)	0.0178(0.0021)	0.0134(0.0024)

known benchmark databases, including some UCI data sets, in two ways: traditional feature-based classification and newly developed dissimilarity-based classification schemes. Our experimental results demonstrate that the classification accuracy of SSMB was improved by employing the proposed learning method. Although we have shown that SSMB can be improved in terms of classification accuracy, many tasks remain. One of them is to improve the classification efficiency by selecting an optimized or nearly optimized number of unlabeled samples for the incremental learning process. The significant data sets best suited for the scheme should be determined. Therefore, the problem of theoretically investigating the experimental results obtained with the proposed SSMB remains to be solved.

REFERENCES

Cesa-Bianchi, N., G. C. Z. L. (2006). Incremental algorithms for hierarchical classification. *Journal of Machine Learning Research*, 7:31–54.

d'Alché Buc, F., G. Y. A. C. (2002). Semi-supervised marginboost. In *Advances in Neural Information Processing Systems*, volume 14, pages 553–560. the MIT press.

Duin, R. P. W., J. P. d. D. P. P. P. E. and Tax, D. M. J. (2004). *PRTools 4: a Matlab Toolbox for Pattern Recognition*. Delft University of Technology, The Netherlands.

Fukunaga, K. (1990). *Introduction to Statistical Pattern Recognition, 2nd*. Academic Press, San Diego, CA.

Mallapragada, P. K., J. R. J. A. K. L. Y. (2009). Semiboost: Boosting for semi-supervised learning. *IEEE Trans. Pattern Anal. and Machine Intell.*, 31(11):2000–2014.

Mason, L., B. J. B. P. L. F. M. (2000). Functional gradient techniques for combining hypotheses. In *Advances in Large Margin Classifiers*. the MIT press.

Pekalska, E. and Duin, R. P. W. (2005). *The Dissimilarity Representation for Pattern Recognition: Foundations and Applications*. World Scientific, Singapore.